

VISIT...

LANZAROTE
Caliente.COM



APPLIED ANALYSIS

BY
CORNELIUS LANCZOS

PRENTICE HALL, Inc., 1956

К. ЛАНЦОШ

ПРАКТИЧЕСКИЕ МЕТОДЫ ПРИКЛАДНОГО АНАЛИЗА

СПРАВОЧНОЕ РУКОВОДСТВО

Перевод с английского
М. З. КАЙНЕРА

под редакцией
А. М. ЛОПШИЦА

ГОСУДАРСТВЕННОЕ ИЗДАТЕЛЬСТВО
ФИЗИКО-МАТЕМАТИЧЕСКОЙ ЛИТЕРАТУРЫ
МОСКВА 1961

АННОТАЦИЯ

Перевод книги известного американского математика Корнелия Ланцоша, одного из виднейших специалистов в области вычислительных методов и их приложений к инженерным проблемам.

Книга состоит из семи глав: I. Алгебраические уравнения. II. Матрицы и проблемы собственных значений. III. Системы многих линейных уравнений. IV. Гармонический анализ. V. Анализ эмпирических данных. VI. Методы квадратур. VII. Степенные разложения.

Книга может быть использована и как справочное пособие: каждый из ее параграфов представляет собой, как правило, отчетливое изложение частного метода, сопровождаемое числовым примером.

Книга предназначена для широкого круга читателей: студентов, математиков-вычислителей, инженеров, применяющих математические методы, работников НИИ, лабораторий и вузов.

К. Ланцош

Практические методы прикладного анализа

Редактор *П. Т. Резниковский*

Техн. редактор *К. Ф. Брудно*

Корректор *З. В. Автонева*

Сдано в набор 8/XII 1960 г. Подписано к печати 18/III 1961 г. Бумага 60×92¹/₁₆. Физ. печ. л. 32,75. Условн. печ. л. 32,75. Уч.-изд. л. 32,09. Тираж 20 000 экз. Т-03117. Цена книги 1 р. 70 к. Заказ № 1240

Государственное издательство физико-математической литературы.
Москва, В-71, Ленинский проспект, 15

Первая Образцовая типография имени А. А. Жданова Московского городского совнархоза.
Москва, Ж-54, Валовая, 28

ОГЛАВЛЕНИЕ

Предисловие редактора перевода	9
Предисловие автора	11
Введение	19
1. Чистая и прикладная математика	19
2. Чистый анализ, практический анализ, численный анализ	20

Г л а в а I

Алгебраические уравнения

1. Историческое введение	23
2. Смежные области	24
3. Кубические уравнения	24
4. Численный пример	26
5. Метод Ньютона	27
6. Численный пример для метода Ньютона	29
7. Схема Горнера	29
8. Техника подвижной полосы	30
9. Остальные корни кубического уравнения	33
10. Подстановка комплексного числа в полином	33
11. Уравнения четвертой степени	36
12. Уравнения высших степеней	39
13. Метод моментов	40
14. Алгебраическое деление двух полиномов	41
15. Степенные суммы и наибольший по модулю корень	43
16. Оценка наибольшего абсолютного значения	47
17. Развертка единичной окружности	49
18. Преобразование обратными радиусами	53
19. Корни, близкие к мнимой оси	56
20. Кратные корни	58
21. Алгебраические уравнения с комплексными коэффициентами	59
22. Анализ устойчивости	60
Литература к главе I	64

Г л а в а II

Матрицы и проблемы собственных значений

1. Исторический обзор	65
2. Векторы и тензоры	67
3. Матрицы как алгебраические объекты	68
4. Анализ собственных значений	73
5. Уравнение Гамильтона — Кели	76
6. Численный пример полного анализа собственных значений	81
7. Алгебраическое доказательство ортогональности собственных векторов	90
8. Геометрическая интерпретация проблемы собственных значений	96

9. Преобразование матрицы к главным осям	105
10. Косоугольная система координат	110
11. Преобразование к главным осям в случае, когда поверхность задана в косоугольной системе	116
12. Инвариантность матричных равенств относительно ортогональных преобразований	125
13. Инвариантность матричных равенств относительно произвольных линейных преобразований	129
14. Коммутативные и некоммутирующие матрицы	132
15. Обращение матриц. Гауссов метод исключения	133
16. Последовательная ортогонализация матрицы	137
17. Обращение треугольной матрицы	144
18. Численный пример последовательной ортогонализации матрицы	147
19. Триангуляризация матрицы	150
20. Обращение комплексной матрицы	151
21. Решение кодиагональных систем	153
22. Обращение матриц путем подразделения на блоки	155
23. Метод возмущений	158
24. Совместность линейных уравнений	163
25. Переопределенность и принцип наименьших квадратов	169
26. Естественная и искусственная косоугольность системы линейных уравнений	173
27. Ортогонализация произвольной линейной системы	175
28. Влияние помех на решение обширных линейных систем	179
Литература к главе II	182

Глава III

Системы многих линейных уравнений

1. Историческое введение	183
2. Операции с матричными полиномами	184
3. p, q -алгоритм	186
4. Полиномы Чебышева	190
5. Спектроскопический анализ собственных значений	192
6. Построение собственных векторов	200
7. Итерационное решение обширных линейных систем	201
8. Остаточное испытание	209
9. Наименьшее собственное значение эрмитовой матрицы	211
10. Собственное значение произвольной матрицы, отличное от наибольшего	214
Литература к главе III	216

Глава IV

Гармонический анализ

1. Исторические замечания	218
2. Основные теоремы	219
3. Квадратичные приближения	222
4. Ортогональность функций Фурье	225
5. Отделение ряда синусов от ряда косинусов	226
6. Дифференцирование ряда Фурье	230
7. Разложение дельта-функции в тригонометрический ряд	232
8. Распространение тригонометрического ряда на неинтегрируемые функции	234
9. Сглаживание колебаний Гиббса с помощью σ -множителей	235
10. Общий характер сглаживания σ -множителями	237
11. Метод тригонометрической интерполяции	238

12. Интерполяция синусами	244
13. Интерполяция косинусами	247
14. Гармонический анализ равноотстоящих данных	249
15. Ошибка тригонометрической интерполяции	251
16. Интерполирование с помощью полиномов Чебышева	254
17. Интеграл Фурье	257
18. Соотношение входа и выхода электрических цепей	264
19. Эмпирическое определение соотношения входа — выхода	268
20. Интерполирование преобразования Фурье	271
21. Интерполяционный анализ фильтра	273
22. Разыскание скрытых периодичностей	276
23. Выделение показательных функций	280
24. Преобразование Лапласа	288
25. Анализ цепей и преобразование Лапласа	290
26. Обращение преобразования Лапласа	292
27. Обращение с помощью полиномов Лежандра	293
28. Обращение с помощью полиномов Чебышева	296
29. Обращение с помощью рядов Фурье	297
30. Обращение с помощью функций Лагерра	300
31. Интерполяция преобразования Лапласа	305
Литература к главе IV	310

Г л а в а V

Анализ эмпирических данных

1. Историческое введение	312
2. Интерполяция с помощью простых разностей	313
3. Интерполяция при помощи центральных разностей	315
4. Дифференцирование табулированной функции	319
5. Неудобства таблицы разностей	319
6. Основной принцип метода наименьших квадратов	321
7. Сглаживание эмпирических данных при помощи четвертых разностей	323
8. Дифференцирование эмпирической функции	327
9. Дифференцирование с помощью интегрирования	330
10. Вторая производная эмпирической функции	332
11. Сглаживание в целом с помощью разложения в ряд Фурье	336
12. Эмпирическое определение граничной частоты γ_0	341
13. Полиномы метода наименьших квадратов	347
14. Полиномиальные интерполяции в целом	349
15. Сходимость полиномиальной интерполяции равноотстоящих данных	355
16. Системы ортогональных функций	360
17. Самосопряженные дифференциальные операторы	364
18. Дифференциальный оператор Штурма — Лиувилля	366
19. Гипергеометрический ряд	368
20. Полиномы Якоби	369
21. Интерполирование ортогональными полиномами	372
Литература к главе V	379

Г л а в а VI

Методы квадратур

1. Исторические замечания	380
2. Квадратура при помощи планиметров	381
3. Правило трапеций	381
4. Правило Симпсона	382
5. Точность формулы Симпсона	385
6. Точность правила трапеций	386

7. Правило трапеций с концевой поправкой	387
8. Численные примеры	390
9. Приближение полиномами высших степеней	393
10. Гауссов метод квадратуры	396
11. Численный пример	401
12. Погрешность квадратуры Гаусса	404
13. Коэффициенты квадратурной формулы с произвольными узлами . . .	406
14. Квадратура Гаусса с округленными узлами	407
15. Применение двукратных корней	409
16. Применения квадратурного метода Гаусса в технике	411
17. Формула Симпсона с концевой поправкой	412
18. Квадратура, содержащая показательные функции	416
19. Квадратура с помощью дифференцирования	417
20. Примеры	422
21. Задачи собственных значений	424
22. Сходимость квадратуры, использующей краевые значения	430
Литература к главе VI	432

Г л а в а VII

Степенные разложения

1. Историческое введение	433
2. Аналитическое разложение с помощью обратных радиусов	435
3. Численный пример	439
4. Сходимость ряда Тейлора	441
5. Жесткие и гибкие разложения	442
6. Разложение по ортогональным полиномам	445
7. Полиномы Чебышева	447
8. Смещенные полиномы Чебышева	449
9. Телескопический сдвиг степенного ряда путем последовательного со- кращения	451
10. Телескопический сдвиг степенного ряда путем пересоставления . . .	453
11. Степенные разложения вне интервала сходимости Тейлора	456
12. τ -метод	457
13. Канонические полиномы	462
14. Примеры приложения τ -метода	467
15. Оценка погрешности τ -метода	483
16. Квадратный корень из комплексного числа	490
17. Обобщение τ -метода. Метод избранных точек	493
Литература к главе VII	496
Приложение. Числовые таблицы	497
Литература	517
Алфавитный указатель	518



ПРЕДИСЛОВИЕ РЕДАКТОРА ПЕРЕВОДА

Среди книг по прикладному анализу, изданных за последние годы, книга Корнелия Ланцоша «Прикладной анализ» резко выделяется не только своим содержанием, но и принципиальными установками автора. Его привлекала не только возможность познакомить широкий круг инженеров и физиков с современными методами численного решения математических задач, возникающих в практической деятельности, но и потребность дать (как сказано на суперобложке английского издания книги) «философское и теоретическое исследование математического аппарата, используемого при этом». И хотя вопросы философского содержания — в прямом смысле этого слова — не рассматриваются в книге Ланцоша, но все же автор уделил внимание рассмотрению некоторых общих принципов, существенных, по его мнению, для развития прикладного анализа за последнее время.

Основную часть книги составляет подробное изложение (сопровожаемое многочисленными, прекрасно подобранными числовыми примерами) разнообразных методов, используемых в настоящее время в вычислительной работе. При этом, однако, автор старается — всюду, где это уместно, — обратить внимание на то, что в промежуток, разделяющий «чистую» и «прикладную» математику, вклинилась в последние десятилетия новая область математики, которую он называет «анализом конечных алгорифмов». Ее задачу он видит в создании и исследовании конечных процессов, приводящих к приближенному решению проблем анализа.

Построение схем, эффективно уменьшающих ошибку (при малом количестве шагов) и дающих возможность оценить ее с достаточной точностью, представляет, по мнению автора, не только практическую ценность, но и самостоятельный научный интерес.

Эта точка зрения отражается на всем изложении материала. Она нисколько не уменьшает удобства использования книги как учебного и справочного издания и несомненно способствует выработке у читателя правильного взгляда на основную сущность задач прикладного анализа; она же приведет его к творческому использованию книги, т. е. к самостоятельному созданию новых вычислительных схем.

Автор книги не ставил себе целью изложить вопросы прикладного анализа во всем их объеме — подробное оглавление, которым

начинается книга, дает полное представление о материале, которым он ограничился.

Он считает, однако, что те семь разделов, из которых состоит книга, содержат «неизбежный арсенал мощного оружия, которое существует теперь для научного исследования инженерных проблем», и полагает, что «вопросы, разобранные в любой главе, встречаются физику и инженеру почти ежедневно».

Добавим к этому, что в свою книгу автор включил большое число методов, которые он самостоятельно развил и практически разработал в период, начавшийся еще в тридцатых годах и продолжающийся до настоящего времени. Характерным для научных устремлений автора является многообразное использование «выдающихся свойств полиномов Чебышева», о которых автор говорит, что они оказали «решающее влияние на мое последующее научное развитие».

Изобретательность, с которой автор использует чебышевский метод аппроксимаций в самых разнообразных вопросах прикладного анализа, действует заражающим образом и несомненно вдохновит вдумчивого читателя на самостоятельное использование этого замечательного средства.

Большое влияние окажет на читателя книги манера изложения автора — свободная, непринужденная, старающаяся «выявить... внутреннюю жизнь математических отношений». Разделяя вместе с автором опасения, что «читатель, которому приходится пробираться в лабиринте «лемм», «следствий» и «теорем», легко сможет заблудиться в формальных деталях в ущерб существенным элементам», переводчик книги и редактор перевода не всегда все же могли согласиться с некоторыми логическими недомолвками, которые встречаются в книге; поэтому они разрешили себе внести в соответствующих случаях изменения (не оговаривая их каждый раз специально), стараясь их редактировать так, чтобы они соответствовали общим установкам автора. В некоторых же местах книги возникла потребность в дополнительных подстрочных примечаниях: в отличие от примечаний автора книги, все редакционные примечания отмечены звездочками.

Указанный автором в конце книги список литературы, которую он рекомендует для параллельного чтения или для более глубокого изучения, дополнен книгами на русском языке и книгами, которые вышли после 1956 г. Это же сделано в списках книг и статей более специального содержания, указанных автором в конце каждой главы.

А. М. Лопшиц

ПРЕДИСЛОВИЕ АВТОРА

В течение многих лет автор занимался исследованиями в тех областях математического анализа, которые интересуют в первую очередь инженеров и физиков. То обстоятельство, что эта область «рабочей математики» не вызывала в течение XIX столетия такого же интереса, как классические разделы анализа, является, пожалуй, плодом исторического недоразумения. Вплоть до эпохи Гаусса и Лежандра «рабочие» методы анализа привлекали к себе пристальное внимание лучших математиков. Блестящее открытие теории пределов изменило положение. С тех пор считалось достаточным указать бесконечный процесс приближения, с помощью которого можно установить правильность определенных аналитических результатов, независимо от того, выполним ли примененный процесс к данной задаче или нет. В результате произошло постепенное разделение «чистой» и «прикладной» математики, так что теперь мы имеем «чистых аналитиков», которые развивают свои идеи в мире чисто теоретических построений, и «численных аналитиков», которые переводят аналитические процессы в машинные операции.

В действительности, однако, существует обширная область, расположенная между этими разделами, которая является не менее аналитической, чем анализ бесконечных процессов, но посвящена совсем другой ветви анализа, а именно анализу конечных алгоритмов. В этой области предметом исследования является анализ и характеристика конечных процессов, которые аппроксимируют решение аналитической задачи. Нахождение приемов, которые при небольшом числе шагов эффективно минимизируют погрешность и дают с достаточной точностью оценку этой погрешности, является делом, имеющим не только практический, но также и научный интерес. Настоящая книга в основном посвящена такого рода проблемам.

Несколько замечаний относительно формы изложения будут, быть может, уместны. Автор не склонен жертвовать строгостью под тем предлогом, что научный работник-практик интересуется только результатами, а не теми более или менее сложными рассуждениями, которые приводят к ним. Математические понятия и утверждения четки и недвусмысленны, и всякое «неряшливое» выражение математической теоремы дисквалифицирует формулировку и вызывает

сомнение в верности утверждаемого результата. Однако можно считать допустимыми формулировки и доказательства теорем при менее широких условиях, чем те, которых добивается чистый аналитик, если при таком ограничении методы и результаты математических исследований могут быть выражены языком не чрезмерно чуждым для научного работника в области физики или техники. Кроме того, автор придерживается того мнения, что математические формулы за своей как бы бесплотной внешностью имеют еще «внутреннюю интимную жизнь». Выявить эту «внутреннюю жизнь» математических отношений порой путем повествовательного отступления не кажется ему профанацией священного ритуала формального анализа, а просто попыткой достичь более полного понимания. Читатель, которому приходится пробираться в лабиринте «лемм», «следствий» и «теорем», легко может заблудиться в формальных деталях в ущерб существенным элементам полученных результатов. Сосредоточив свое внимание на основных моментах, он выигрывает в глубине, хотя и может потерять в подробностях. Эта потеря, однако, не серьезна, так как читатель, владеющий элементарными средствами алгебры и анализа, легко может восполнить недостающие подробности. Хорошо известно из опыта, что единственно приятным и полезным путем изучения математики является метод самостоятельного дополнения подробностей. Эта дополнительная работа, как надеется автор, возбудит воображение читателя и, может быть, будет стимулировать изучение и дальнейшие исследования как на учебном, так и на исследовательском уровне.

Само собой разумеется, что книга такого рода не может исчерпать предмет полностью, не разросшись чрезмерно. Обширную область задач на краевые значения вместе с теорией интегральных уравнений пришлось опустить за недостатком места. Но не будет, пожалуй, преувеличением сказать, что вопросы, разобранные в любой главе, встречаются физику и инженеру почти ежедневно. Ниже мы приводим краткое содержание каждой главы.

Глава I. Алгебраические уравнения. Отыскивать корни алгебраического уравнения часто приходится в задачах теории колебаний и неустановившихся вибраций, а также в задачах статического и динамического равновесия. В гл. I рассматриваются некоторые полезные приемы вычислительной техники, использующие метод «подвижной полосы». Центральную роль в гл. I играет метод Бернулли со всеми своими ответвлениями, но интерес представляет также метод развертки единичной окружности для отделения комплексных корней почти равного модуля и метод обратных радиусов.

Глава II. Матрицы и проблемы собственных значений. Эта глава посвящена систематическому изложению свойств матриц, причем в основном внимание обращено на те особенности, которые чаще всего встречаются в технических исследованиях.

Глава III. Системы многих линейных уравнений. Появление электронных вычислительных машин выдвинуло на первый план

итерационные методы при решении сложных задач на краевые значения и задач вибрации. Это привело одновременно к исследованию операций с матричными полиномами. Хотя общий случай комплексных собственных значений не мог быть включен в изложение, «спектроскопический метод» нахождения вещественных собственных значений матриц высокого порядка и соответствующий метод решения системы линейных уравнений с большим числом неизвестных имеют столь большое применение (и столь естественно связаны со следующей главой — о гармоническом анализе), что их разбора едва ли можно было избежать. Изложение метода возмущений дает по крайней мере частичное решение задачи нахождения комплексных собственных значений; здесь показывается, как можно получить любое комплексное собственное значение и собственный вектор, если мы имеем для начала достаточно хорошее первое приближение искомого собственного значения.

Глава IV. Гармонический анализ. Большие размеры этой главы могут быть оправданы чрезвычайно важным значением ряда Фурье и его следствий, интеграла Фурье и преобразования Лапласа для всех разделов анализа. Можно, пожалуй, перефразировать знаменитое выражение Виктора Гюго о том, что если бы ему предложили уничтожить всю литературу, оставив только одну книгу, он сохранил бы книгу Иова. Подобно этому, если бы нам предложили выбросить все математические открытия, кроме одного, мы едва ли бы не оставили ряд Фурье. Этот ряд оказал наиболее глубокое влияние на все развитие анализа как в его теоретическом, так и практическом аспектах. Кроме того, его связь с другими частями анализа столь тесна, что если мы сказали бы «ряд Фурье со всеми его следствиями», то значительная часть нашего классического анализа была бы сохранена.

Для приложений к технике наиболее важным фактом является ортогональность функций Фурье для случая равноотстоящих данных. В соответствии с этим в гл. IV прежде всего рассматриваются интерполяционные аспекты ряда Фурье, его гибкость в представлении функций по эмпирически полученным равноотстоящим данным. Для приведения ряда к практически пригодному виду применяется искусственный прием — вычитание линейного остатка. Этот прием приводит к нулю оба краевых значения и дает возможность использовать ряд из одних синусов. Остающиеся при этом колебания Гиббса, обусловленные разрывностью второй производной, слишком малы чтобы причинить какие-нибудь неудобства.

Другой искусственный прием — применение « σ -множителей» — противодействует колебаниям Гиббса, вызывающим расхождение ряда при его дифференцировании. Тот же метод сглаживает неприятные колебания Гиббса в окрестности точки разрыва и при представлении дельта-функции. В анализе электрических цепей методы преобразований Фурье и Лапласа получили огромный толчок за последние несколько лет. Эти преобразования представляют собой не только

теоретические методы для доказательства некоторых основных теорем, но и играют фундаментальную роль при практическом построении соотношения входа — выхода линейных цепей. Интерполяционное решение задачи фильтров и различные методы обращения преобразования Лапласа рассматриваются как проблемы прикладного анализа, имеющие весьма существенное техническое значение. Наконец излагается часто встречающаяся задача «разыскания скрытых периодичностей»; при этом развивается численная схема, при которой достигается большая независимость от различных частот и тем самым большая разрешающая сила и большая точность, чем при обычных схемах.

Глава V. Анализ эмпирических данных. Задача сглаживания эмпирических данных и задача нахождения первой и даже второй производной эмпирически заданной функции постоянно встречаются при изучении проблемы движения, а также при построении эмпирических кривых.

В главе V рассматриваются два метода сглаживания: сглаживание в малом и сглаживание в целом. В первом случае применяются локальные квадратичные параболы, которые приводят к определенному взвешенному среднему каждого наблюдения вместе с несколькими левыми и правыми смежными данными. Во втором случае производится разложение в ряд Фурье, и сглаживание достигается просто обрыванием ряда в надлежаще выбранном месте. Последний метод дает вместе с тем аналитическое выражение, представляющее все наши данные и интерполирующее их в любой точке интервала. Иногда желательно получить хорошее полиномиальное приближение; для этого случая описывается метод, преобразующий ряд Фурье в очень хорошо сходящийся ряд полиномов. Мы, таким образом, избегаем недостатков равноотстоящего интерполирования Лагранжа и получаем полином, хорошо согласующийся с нашими данными при небольших и практически равномерно распределенных погрешностях.

Глава VI. Методы квадратур. Еще на заре математической науки задача интегрирования приковывала внимание ученых. Каждый период развития науки вносил свою долю в познание проблемы квадратур. Архимед, впервые рассматривая интегрирование как точный предельный процесс, пользовался для целей квадратуры только трапециями с постоянно убывающими сторонами; в последующие века совершенствовали технику интерполирования, оперируя полиномами более высоких порядков.

Глава VI дает обзор различных методов квадратур. Квадратура Гаусса занимает среди них преобладающее место, как наиболее совершенный из всех квадратурных методов. Описывается полезное видоизменение этого метода, в котором устранена необходимость интерполяции в иррациональных точках. Кроме того, основная идея метода Гаусса переносится на случай, когда мы располагаем лишь краевыми значениями, а именно значениями функции и ее производных до определенного порядка на обоих концах интервала. Получающаяся квад-

ратурная формула имеет сильную сходимость и может быть применена для решения задач на краевые значения и задач собственных значений, связанных с обыкновенными дифференциальными уравнениями. Эта формула иллюстрируется несколькими численными примерами.

Глава VII. Степенные разложения. Метод представления функций с помощью полиномов известен давно. Однако представление функции на заданном интервале комбинацией небольшого количества степенных функций с небольшой погрешностью выходит за пределы разложения в ряд Тейлора и требует использования теории систем ортогональных функций. Существует, в частности, один специальный класс полиномов, «полиномы Чебышева», который обеспечивает более сильную сходимость, чем любой другой класс полиномов. На самом деле мы опять возвращаемся к ряду Фурье, так как разложение по полиномам Чебышева в действительности есть не что иное, как разложение по косинусам для новой переменной.

Мы можем с успехом использовать эти полиномы для решения обыкновенных линейных дифференциальных уравнений, коэффициенты которых суть рациональные функции от x . Так как большинство основных трансцендентных функций, встречающихся в математической физике, могут быть определены с помощью таких дифференциальных уравнений, то мы таким образом получаем быстро сходящиеся разложения для многих различных функций. Мы с самого начала ограничиваем наш ряд конечным полиномом заданного порядка, что приводит к необходимости введения поправочного члена в правой части заданного дифференциального уравнения. Этот поправочный член принимается пропорциональным полиному Чебышева надлежащего порядка. При этом мы получаем простые рекуррентные соотношения, из которых могут быть определены коэффициенты аппроксимирующего полинома. Иногда необходимо повторить несколько раз этот процесс, тогда окончательный результат получается в виде линейной суперпозиции составляющих полиномов. Этот « τ -метод» делает степенное разложение максимально эффективным и оказывается очень полезным и в тех случаях, когда мы имеем в виду аппроксимирование элементарной функции в операторных целях, и в случае аппроксимирования трансцендентной функции в целях ее вычисления. Уменьшение погрешности, которое достигается в τ -методе по сравнению с разложением в ряд Тейлора (если таковое вообще существует), всегда весьма значительно.

В конце книги приведен список книг, которые автор рекомендует для параллельного чтения или для более глубокого изучения. В него вошли только такие книги, содержание которых либо подобно содержанию настоящей книги, либо включает в себе относящийся к нему материал. Если какие-либо важные источники не вошли в этот список, то это сделано не намеренно. Книги Куранта — Гильберта и Уиттекера — Ватсона представляют собой незаменимый источник информации. В области численного анализа основными являются книги Милна,

Скарборо и Хартри. Книги и статьи более специального содержания указаны в конце каждой главы. Ссылки, отмеченные знаком {}, относятся к книгам, указанным в этой общей библиографии; ссылки, отмеченные знаком [], относятся к книгам и статьям, указанным в конце глав.

Настоящая книга создавалась в течение ряда лет научных размышлений. В эти годы мне посчастливилось вести беседы с коллегами и друзьями, послужившие основой моей работы. Число лиц, которые мне в этом помогали, столь велико, что перечислить их невозможно. Будет более уместным назвать те учреждения, с которыми я был связан за время созревания моих мыслей.

1. Во время моего пребывания в Университете Purdue (1931—1945) доктор Маршалл (W. Marshall) из математического факультета и доктор Ларк-Горовиц (K. Lark-Horowitz) из физического факультета организовали курс лекций по «Приближенным методам анализа», здесь я впервые столкнулся с задачами аппроксимирования. В эти же годы мне раскрылись выдающиеся свойства полиномов Чебышева, оказавших решающее влияние на мое последующее научное развитие.

2. Во время второй мировой войны я провел 1943—1944 гг. при Бюро математических таблиц в Нью-Йорке, возглавляемом директором доктором А. Н. Лоуэном (A. N. Lowan). Общение с выдающимся штатом численных аналитиков было весьма благоприятным.

3. В течение 1944—1945 гг. я прочитал два курса лекций: «Технические приложения быстро сходящихся рядов» и «Курс математических приближений», организованные Бюро физических исследований при Авиационной компании Боинг в Сиэтле, Вашингтон. Отлично изданные mimeографические записки этих лекций, составленные при умелой помощи Мариуса Кона (Marius Cohn), образовали основное ядро настоящей книги.

В 1946 г. доктор Ч. К. Стедмен (C. K. Stedman), нач. отдела физических исследований, привлек меня в состав отдела в качестве математического консультанта и инженера-исследователя. То значение, которое имели для меня повседневные беседы с группой выдающихся физиков и инженеров-электриков, нельзя переоценить.

4. По приглашению доктора Ж. Г. Кэртисса (J. H. Curtiss), тогда возглавлявшего Национальное Бюро Стандартов в Вашингтоне, я работал в 1949 г. во вновь основанном Институте численного анализа при Калифорнийском университете в Лос-Анжелосе. Под руководством Кэртисса институт обеспечил научную атмосферу и коллегиальное общение на таком уровне, которому мало равного где-либо в мире. То великодушное, с которым институт поддерживал мои научные предложения, навсегда останется в моей памяти.

5. По приглашению Дублинского института высших исследований я провел памятный год (1952/53) в Дублине, Ирландия, в ежедневном контакте с директором института, профессором Е. Шредингером и профессором Ж. Синджем (J. L. Synge). В течение этого года были

получены квадратурная формула § 22 гл. VI и ее приложения к задачам собственных значений.

6. В течение зимы 1953/54 г. я был зачислен в качестве штатного сотрудника отдела табулирования Ч. Ф. Дэвиса (Davis), численного аналитика. Прочитанный там курс лекций, который посещался группой крупных инженеров, привел к полезным дружеским связям и нашел отражение во второй и третьей главе этой книги. В это время был развит «спектроскопический анализ собственных значений».

Наряду с тем исключительным содействием, которое мне посчастливилось встретить в научной работе, большая помощь была мне оказана при составлении численной документации математических теорий. За выполнение численных примеров книги я благодарю прежде всего мисс Мери Эллен Рассел, исследователя-ассистента Бюро физических исследований Авиационной компании Боинг; таблицы, помещенные в приложении, составлены главным образом мисс Лилиан Фортселл из Института численного анализа Национального Бюро Стандартов, Лос-Анжелос.

Корнелий Ланцош

Дублинский институт
высших исследований,
Дублин, Ирландия

ВВЕДЕНИЕ

1. Чистая и прикладная математика

История математики показывает, что интерес к формальным математическим операциям почти всегда был связан с желанием получить правильную картину физического мира. Постулаты и операции анализа выбраны не *произвольно*, а так, что они отображают геометрический порядок вещей в абстрактной области чисел. Формальные операции анализа являются, таким образом, лишь одним звеном в нашем стремлении раскрыть функциональную закономерность, присущую физическому миру. В целом этот процесс содержит три фазы:

1. Данное физическое соотношение переносится в область чисел.
2. С помощью чисто формальных операций над этими числами получают определенные математические результаты.
3. Эти результаты переносятся обратно в мир физической реальности.

С развитием математики вторая фаза этого процесса соотнесения в конце концов становится независимой замкнутой в себе научной дисциплиной. Мы можем углубиться в формальные аналитические операции, не задаваясь вопросом, имеет ли поставленная проблема соответствующий аналог в природе. Точно так же мы можем остановиться на аналитическом результате, полученном во второй фазе, не интерпретируя его обратно в физические закономерности вещей.

XIX столетие изобрело термины: «чистая» и «прикладная» математика для характеристики этих двух положений; терминология эта далека от того, чтобы быть точной и удовлетворительной. Термин «чистая математика», выражающий выделение формальных математических операций, имеет то привходящее значение, что этот вид умственной деятельности является более «чистым», более «искусством для искусства» и, таким образом, более «идеалистическим», чем другой вид деятельности, который теснее связан с закономерностями, присущими физическому миру. Это дополнительное значение едва ли основательно, и оно могло быть устранено при более точной терминологии. Однако мы часто продолжаем употреблять слова, даже если они были выдуманы неудачно, наваяны ассоциациями, которые не получили научного оправдания. Выражения «отрицательные числа» и

«мнимые числа» возникли тогда, когда еще плохо понимали истинный смысл этих понятий. Эти названия, однако, сохранились и вызывают недоразумения в тех случаях, когда уделяется недостаточное внимание условности в определении этих слов. Такое же положение существует и с неправильным наименованием «чистой» и «прикладной» математики.

2. Чистый анализ, практический анализ, численный анализ

Существуют, однако, несколько иные обстоятельства, также вызывающие необходимость характерной терминологии, при которой опять вводятся в употребление слова «чистый» и «практический» без надлежащего оправдания. Уже математикам древней Греции IV и III веков до н. э. был известен тот важный факт, что некоторые математические проблемы требуют построения *бесконечной последовательности никогда не заканчивающихся приближений*, которые становятся все более близкими к искомому результату, никогда не достигая его в точности. В своем фундаментальном трактате о круге Архимед открыл, что длина окружности не только не может быть точно *вычислена*, но даже не может быть *определена* точно. Точные алгебраические методы были поэтому заменены другого рода подходом, который греки называли «методом исчерпывания»; математики XIX века приняли название «метода пределов». Применяя этот метод пределов, часто приводящий к разложению в бесконечный ряд, мы стремимся не *получить* искомую величину, а только *приблизиться* к ней с ошибкой, которая может быть сделана сколько угодно малой. «Абсолютная точность», таким образом, заменяется «сколь угодно большой точностью». Точность первого рода, которую можно достичь в алгебре, но *нельзя* достичь в высшей математике, означает «нулевую ошибку». Второго рода точность, типичная для методов высшей математики, означает «конечную, но сколь угодно малую ошибку».

Широкие перспективы, открытые теорией пределов Коши и Гаусса, привели к подавляющему господству *методов бесконечного приближения*. При этих методах число выполненных шагов не имеет никакого значения. Нас не интересует, что случится при *конечном* числе шагов. Нам достаточно знать, что случится *в конце концов*, когда число шагов возрастет безгранично.

Меньше внимания уделялось другого рода методу приближения, который логически столь же возможен. Мы можем пожелать узнать, какую ошибку мы допускаем, если применим некоторый процесс определенное *конечное* число раз. В частности, нас может интересовать установление такого аппроксимирующего процесса, при котором получается искомый результат с наименьшей ошибкой, при заданном числе шагов. Здесь точность не имеет характера ни абсолютной, ни произвольно большой точности. Мы примиряемся с тем фактом, что

ошибка совершается. Но мы хотели бы иметь средства для *оценки* этой ошибки. Кроме того, мы хотели бы исследовать, какие приемы можно применить для эффективного *уменьшения* этой ошибки.

Эта ветвь анализа, которой исторически уделялось гораздо меньше внимания, чем описанному ранее предельному процессу, не получила надлежащего названия. Однако в то время как классический или «чистый» анализ занимается процессами бесконечного аппроксимирования, другая ветвь анализа обозначается иногда термином «практический анализ». Это название опирается на тот факт, что физика и техника при определенном процессе аппроксимирования не интересуется уменьшением ошибки до нуля или до произвольно малой величины, так как наши наблюдения имеют во всяком случае ограниченную точность. Поэтому мы удовлетворяемся ответом, который «теоретически» несовершенен, но «практически» приемлем. Кроме того, достижение теоретически совершенного ответа может потребовать использования приемов слишком громоздких для практического осуществления. Поэтому мы в практическом анализе занимаемся конечными процессами, которые могут быть численно проведены и дают при относительно небольшом числе шагов такую точность, которая считается достаточной для определенных задач физики и техники.

Противопоставление слов «чистый» и «практический» снова здесь сопровождается неправильной оценкой. Мы склонны думать, что только процессы «чистого» анализа имеют ясно выраженный математический интерес, тогда как процессы «практического» анализа строятся лишь ввиду нужд «практиков», т. е. инженеров и физиков. Однако, согласившись с этим, мы приписали бы абсолютное значение случайному обстоятельству исторического развития. Рассмотрим в качестве примера разложение в бесконечный ряд числа π , который для получения точности в 1% потребует вычисления суммы 100 членов ряда. Если другое разложение «практического анализа» дает ту же точность при вычислении лишь пяти членов, то мы вправе принять, что второе разложение даже с чисто аналитической точки зрения является более подходящим, чем первый весьма слабо сходящийся ряд. Более сильную сходимость второго ряда можно рассматривать, как «практическое» достижение, но ее можно считать также и результатом, имеющим чисто логический интерес.

Едва ли можно отрицать, что мы здесь имеем пред собой ветвь математического анализа, которая заслуживает более подходящего названия, чем слово «практический», со своим утилитарным оттенком. Настоящая книга делает шаг в этом направлении, вводя технический термин для обозначения этого класса аналитических методов. Греческое слово «парэктический» (с корнями «пара»-почти, как бы и «эк» — вне) означает «приблизительный». Поэтому термин «парэктический анализ» можно считать подходящим для обозначения того, что нам нужно не точное, а лишь «приближенное» определение некоторой величины. Мы можем, следовательно, говорить о «парэкти-

ческих» методах, «парэктических» разложениях, о «парэктической» точке зрения, в отличие от соответствующих методов, разложений и точки зрения «чистого анализа», который стремится к достижению произвольной точности с помощью бесконечных процессов. Автор считает, что такая терминология вполне вознаграждается возможностью кратко обозначить то, что при обычной терминологии может быть выражено лишь при помощи многословных объяснений¹⁾).

За последние несколько лет вошло в обращение новое слово: «численный анализ». Этот термин хорошо подходит для обозначения известных аспектов математического анализа, которые занимаются переводом математических процессов в операции над числами. Основным здесь является та легкость, с которой некоторые аналитические методы могут быть проведены при замещении математических величин явными числами; эта ветвь анализа изучает ошибки округления, обусловленные не приближенным характером парэктических процессов, а приближенным характером арифметических операций умножения и деления, когда мы ограничиваемся конечным числом десятичных знаков.

Таким образом, мы приходим к заключению, что чистый анализ, парэктический анализ и численный анализ представляют собой три описанные выше фазы математических исследований, каждая из которых имеет свою собственную область интересов и пользуется свойственными каждой из них характеристическими методами. Настоящая книга посвящена главным образом области «парэктического анализа», однако в ней не упускаются из виду общие задачи чистого анализа и более вычислительные аспекты численного анализа.

¹⁾ Только в этом единственном случае настоящая книга отклоняется от обычной математической терминологии (либо отсутствия терминологии). Автор хорошо понимает, что образование новых слов является привилегией гораздо более выдающихся умов, но в данном случае, очевидно, существует крайняя необходимость нового термина, и если предложение автора не будет принято, то оно по крайней мере обратит внимание на существование этой проблемы.

ГЛАВА I

АЛГЕБРАИЧЕСКИЕ УРАВНЕНИЯ

1. Историческое введение

Алгебраические уравнения первой и второй степени вызывали интерес ученых с древнейших времен. В то время как древние египтяне решали главным образом уравнения первой степени, древние вавилоняне (около 2000 лет до н. э.) были уже знакомы с решением квадратного уравнения и составили таблицы для решения кубических уравнений путем приведения общего кубического многочлена к нормальному виду.

Индийцы развили последовательную алгебраическую теорию уравнений первой и второй степени (VII век). Обычный метод решения квадратного уравнения путем дополнения трехчлена до полного квадрата является их изобретением. Индийцы были знакомы с формальной точкой зрения на математические операции, и они не боялись употреблять отрицательные числа, рассматривая их как «долги». Ясное понимание природы мнимых и комплексных чисел пришло значительно позже, во времена Эйлера (XVIII век).

Решение кубического уравнения было найдено итальянцем Тарталья (начало XVI века); ученик Кардана Феррари получил несколькими годами позже решение уравнений четвертой степени. Существенно отличный характер уравнений пятой и высших степеней был ясно осознан Лагранжем (конец XVIII века), но первое точное доказательство того, что общее уравнение пятой и более высокой степени не может быть решено чисто алгебраическими средствами — принадлежит норвежцу Абелю (1824), тогда как спустя несколько лет (1832) француз Галуа дал общее теоретико-групповое обоснование всей проблемы.

«Основная теорема алгебры» устанавливает, что каждое алгебраическое уравнение имеет по крайней мере один корень в области комплексных чисел. Если это доказано, то мы можем немедленно вывести (путем последовательного деления на множители, соответствующие каждому корню), что каждый полином степени n может быть разложен на n множителей. Первое строгое доказательство основной теоремы алгебры было дано двадцатидвухлетним

Гауссом (1799). Позднее теория функций комплексного переменного, созданная Коши, привела к более глубокому взгляду на природу корней алгебраического уравнения и позволила дать более простое доказательство основной теоремы.

Существование в точности n , вообще говоря, комплексных корней алгебраического уравнения n -й степени отнюдь не противоречит неразрешимости алгебраического уравнения пятой и более высокой степени средствами алгебры. Последнее утверждение означает, что корни общего алгебраического уравнения степени выше четвертой не могут быть получены путем чисто алгебраических операций над коэффициентами (т. е. сложения, вычитания, умножения, деления, возведения в степень и извлечения корня). Однако с помощью этих операций можно *аппроксимировать* корни с любой степенью точности.

2. Смежные области

а) Проблема решения алгебраического уравнения n -й степени тесно связана с теорией колебаний около положения равновесия. Частоты колебаний механической системы, вернее квадраты частот, являются «характеристическими корнями» или «собственными значениями» матрицы, получающимися при решении «характеристического уравнения» матрицы, а это уравнение является алгебраическим уравнением n -й степени.

б) В электротехнике характеристика электрической цепи является всегда линейной суперпозицией показательных функций. Показатели этих функций получаются как корни полинома, который может быть построен, если заданы элементы цепи и ее схема.

с) Сложные алгебраические и геометрические соотношения путем исключения часто приводятся к алгебраическому уравнению второй или более высокой степени от одного из неизвестных.

3. Кубические уравнения

Уравнения третьей и четвертой степени еще могут быть решены с помощью алгебраических формул. Однако численные расчеты по этим формулам обычно весьма запутаны и трудоемки. Поэтому мы предпочитаем менее громоздкие методы, которые дают лишь приближенные, но все же настолько близкие к истинным значения корней, что можно применить методы, проводящие к дальнейшему уточнению.

Процесс решения кубического уравнения (с вещественными коэффициентами) особенно прост, так как один из корней всегда должен быть вещественным. Найдя этот корень, мы сразу же получаем другие два корня, решая квадратное уравнение.

Общее кубическое уравнение можно привести к виду

$$\xi^3 + a\xi^2 + b\xi - c = 0, \quad c > 0. \quad (1-3.1)$$

Коэффициент при ξ^3 всегда может быть приведен к единице, так как мы можем разделить обе части уравнения на коэффициент при высшем члене. Кроме того, свободный член можно всегда сделать отрицательным, ибо если он первоначально положителен, мы полагаем $\xi_1 = -\xi$ и оперируем с этим новым неизвестным ξ_1 .

Удобно, далее, ввести новый масштабный множитель, который приводит (нормирует) свободный член к -1 . Положим

$$x = \alpha \xi, \quad a_1 = \alpha a, \quad b_1 = \alpha^2 b, \quad c_1 = \alpha^3 c \quad (1-3.2)$$

и запишем вновь полученное уравнение в виде

$$f(x) = x^3 + a_1 x^2 + b_1 x - c_1 = 0. \quad (1-3.3)$$

Если принять

$$\alpha = \frac{1}{\sqrt[3]{c}}, \quad (1-3.4)$$

то получим

$$c_1 = 1. \quad (1-3.5)$$

Учитывая теперь, что $f(0)$ отрицательно, а $f(\infty)$ положительно, приходим к выводу, что между $x=0$ и $x=\infty$ должен быть по меньшей мере один вещественный корень. Положим $x=1$ и вычислим $f(1)$. Если $f(1)$ положительно, то должен существовать корень, лежащий между 0 и 1; если $f(1)$ отрицательно — корень, лежащий между 1 и ∞ . Кроме того, из равенства

$$x_1 \cdot x_2 \cdot x_3 = 1 \quad (1-3.6)$$

следует, что все *три* корня не могут располагаться между 0 и 1 или между 1 и ∞ . Следовательно, если $f(1) > 0$, то в интервале $[0, 1]$ должен лежать один и только один вещественный корень, тогда как при $f(1) < 0$ один и только один вещественный корень должен быть в интервале $[1, \infty]$ *). Последний интервал может быть переведен в интервал $[1, 0]$ с помощью преобразования

$$\bar{x} = \frac{1}{x}, \quad (1-3.7)$$

которое приводит только к тому, что коэффициенты уравнения меняют свой порядок следования:

$$-c\bar{x}^3 + b_1\bar{x}^2 + a_1\bar{x} + 1 = 0. \quad (1-3.8)$$

Таким образом, мы свели нашу задачу к новой: найти вещественный корень кубического уравнения в интервале $[0, 1]$. Эту задачу

*) Это следует из известной теоремы алгебры о том, что если левая часть алгебраического уравнения $f(x)=0$ с вещественными коэффициентами на границах некоторого интервала $[a, b]$ имеет разные знаки, то уравнение имеет внутри интервала *нечетное* число вещественных корней.

мы решим с достаточно хорошим приближением, используя замечательные свойства полиномов Чебышева (ср. VII, 9). При помощи этих полиномов можно выразить с небольшой ошибкой какую-либо степень x через более низкие степени. В частности, третий полином Чебышева

$$T_3^*(x) = 32x^3 - 48x^2 + 18x - 1, \quad (1-3.9)$$

построенный для интервала $[0, 1]$, дает

$$x^3 = \frac{48x^2 - 18x + 1}{32} = 1,5x^2 - 0,5625x + 0,03125 \quad (1-3.10)$$

с максимальной ошибкой в $\pm \frac{1}{32}$. Первоначальное кубическое уравнение можно, таким образом, свести к квадратному с ошибкой, не превышающей 3% .

Остается только решить это квадратное уравнение и взять тот его корень, который лежит между 0 и 1.

4. Численный пример

На практике нет нужды находить α с большой точностью; можно ограничиться двумя значащими цифрами. Рассмотрим решение следующего кубического уравнения:

$$\xi^3 + 1,2\xi^2 - 17,0\xi - 70 = 0. \quad (1-4.1)$$

Соответствующие таблицы дают для $\sqrt[3]{70}$ значение 4,1212. Обратное значение корня $\alpha = 0,2426$. Для удобства принимаем

$$\alpha = 0,25 \quad (1-4.2)$$

и получаем

$$f(x) = x^3 + 0,3x^2 - 1,0625x - 1,09375 = 0. \quad (1-4.3)$$

При $x = 1$ $f(1) = -0,856$ еще отрицательно. Корень, следовательно, лежит между $x = 1$ и $x = \infty$. Обращаем эту область, полагая

$$\bar{x} = \frac{1}{x}. \quad (1-4.4)$$

Тогда

$$1,09375\bar{x}^3 + 1,0625\bar{x}^2 - 0,3\bar{x} - 1 = 0. \quad (1-4.5)$$

Подстановка (3.10) приводит это уравнение к квадратному уравнению

$$2,703\bar{x}^2 - 0,915\bar{x} - 0,966 = 0, \quad (1-4.6)$$

решение которого дает

$$\bar{x} = \frac{0,915 \pm 3,370}{5,406}. \quad (1-4.7)$$

Знак «минус» приводит к ненужному для нас результату, так как соответствующее значение $\bar{x}$ лежит вне рассматриваемого интервала; знак «плюс» дает

$$\bar{x} = 0,79 \quad (1-4.8)$$

и, следовательно,

$$x = \frac{1}{0,79} = 1,27 \quad \text{и} \quad \xi = \frac{1,27}{0,25} = 5,08. \quad (1-4.9)$$

Подстановка в первоначальное уравнение показывает, что левая часть принимает значение 5,692, что в сравнении со свободным членом 70 составляет ошибку в 8% *).

Действия с большими корнями не выгодны, так как они приводят к увеличению погрешности. Существенно поэтому, что мы можем всегда ограничиться вычислением корней, абсолютное значение которых меньше 1, ибо если оно больше единицы, то преобразование $\bar{x} = \frac{1}{x}$ (которое только обращает порядок следования коэффициентов полинома) заменяет корень на обратный. Следовательно, в нашем примере мы предпочтем подстановку числа, обратного указанному в формуле (9)¹⁾, т. е.

$$\bar{\xi} = \frac{1}{5,08} = 0,197, \quad (1-4.10)$$

в обращенное уравнение

$$70\bar{\xi}^3 + 17\bar{\xi}^2 - 1,2\bar{\xi} - 1 = 0. \quad (1-4.11)$$

Остаток теперь равен $-0,0395$, что составляет ошибку в 4% сравнительно со свободным членом 1.

5. Метод Ньютона

Имея приближенное значение $x = x_0$ корня алгебраического уравнения, мы можем улучшить это значение методом, известным под названием «метод Ньютона». Положим

$$x = x_0 + h \quad (1-5.1)$$

*) Сравнение значения функции со свободным членом имеет следующий смысл: если полученное значение окажется меньше той погрешности, с которой задан свободный член, то значение ξ можно считать корнем, найденным с достаточной точностью. Следует отметить, что при этом не учитываются погрешности остальных коэффициентов многочлена, так что определяемая указанным выше условием точность корня является достаточной, но не необходимой.

¹⁾ Формулы, находящиеся в текущем параграфе, обозначаются при ссылках лишь числом, следующим за точкой. Поэтому (9) означает (1-4.9), так как эту формулу следует искать в текущем параграфе. Формула, содержащаяся в текущей главе, но не в текущем параграфе, обозначается при ссылках номером параграфа и формулы. Следовательно, (3.9) означает формулу (1-3.9)

и разложим $f(x_0 + h)$ в ряд по степеням h :

$$f(x_0 + h) = f(x_0) + hf'(x_0) + \frac{h^2}{2}f''(x_0) + \dots \quad (1-5.2)$$

При малом h члены со старшими степенями h быстро убывают. Если пренебречь всеми членами ряда, кроме первых двух, то получим решение уравнения

$$f(x) = f(x_0 + h) = 0 \quad (1-5.3)$$

с хорошим приближением, приняв

$$h_0 = -\frac{f(x_0)}{f'(x_0)}. \quad (1-5.4)$$

Теперь мы можем рассматривать

$$x_1 = x_0 + h_0 \quad (1-5.5)$$

как новое приближение и применить метод уже к x_1 . Поэтому

$$h_1 = -\frac{f(x_1)}{f'(x_1)} \quad (1-5.6)$$

в сочетании с

$$x_2 = x_0 + h_0 + h_1 \quad (1-5.7)$$

дает еще более близкое значение корня.

Таким образом, мы получаем следующую итерационную схему:

$$h_n = -\frac{f(x_n)}{f'(x_n)}, \quad (1-5.8)$$

$$x_{n+1} = x_0 + h_0 + \dots + h_n. \quad (1-5.9)$$

Эта схема дает быструю сходимость к x , если x_0 есть достаточно близкое первое приближение.

Схема Ньютона применима не только к алгебраическим, но и к трансцендентным уравнениям.

Можно ускорить сходимость метода Ньютона, если остановиться в ряде (2) лишь после третьего члена, рассматривая член второго порядка как малую поправку к члену первого порядка. Тогда

$$f(x_0) + h \left(f'(x_0) + \frac{h}{2} f''(x_0) \right) = 0, \quad (1-5.10)$$

и мы получаем решение этого уравнения в виде

$$h = -\frac{f(x_0)}{f'(x_0) + \frac{h}{2} f''(x_0)}. \quad (1-5.11)$$

Замена h в знаменателе первым приближением (4) приводит к формуле, которую лучше всего запомнить в виде

$$\frac{1}{h_0} = -\frac{f'(x_0)}{f(x_0)} + \frac{1}{2} \frac{f''(x_0)}{f'(x_0)}. \quad (1-5.12)$$

6. Численный пример для метода Ньютона

В § 4¹⁾ было рассмотрено кубическое уравнение

$$f(x) = 1,09375x^3 + 1,0625x^2 - 0,3x - 1 = 0 \quad (1-6.1)$$

и получено приближенное значение корня

$$x_0 = 0,79. \quad (1-6.2)$$

Подставляя это значение в $f(x)$, а также в

$$f'(x) = 3,28125x^2 + 2,125x - 0,3 \quad (1-6.3)$$

и в

$$\frac{1}{2}f''(x) = 3,28125x + 1,0625, \quad (1-6.4)$$

получим

$$\begin{aligned} f(x_0) &= -0,034631, & f'(x_0) &= 3,42658, \\ \frac{1}{2}f''(x_0) &= 3,65468. \end{aligned} \quad (1-6.5)$$

Подстановка в формулу (5.12) дает

$$\frac{1}{h} = 98,945453 + 1,066568 = 100,012021, \quad (1-6.6)$$

$$h = 0,009998798, \quad (1-6.7)$$

$$x_1 = x_0 + h = 0,7999988. \quad (1-6.8)$$

Если подставить это новое значение x_1 в $f(x)$, то получим

$$f(x_1) = -0,00000418. \quad (1-6.9)$$

На этом мы можем остановиться, так как ошибка составляет только четыре единицы шестого десятичного разряда; коэффициенты алгебраического уравнения редко даются с точностью большей, чем до пятого десятичного знака.

7. Схема Горнера

Вычисление результата подстановки некоторого числа в полином выполняется достаточно просто, если это число вещественное, а полином имеет низкую степень. Однако, имея в виду последующие случаи, когда нам придется либо рассматривать полиномы более высоких степеней, либо подставлять в полином комплексные числа, мы разберем более изящную схему, так называемую «схему Горнера». При этой схеме значения $f(x_0)$, $f'(x_0)$, ... получаются в процессе последовательных алгебраических делений *).

¹⁾ Повсюду в книге знак § относится к параграфу текущей главы.

*) Автор употребляет термин «синтетическое» деление, которое в русской литературе мало употребительно. Мы здесь и далее переводим этот термин словами «алгебраическое» деление.

Рассмотрим алгебраическое деление

$$\frac{f(x) - f(x_0)}{x - x_0} = f_1(x). \quad (1-7.1)$$

Если $f(x)$ — полином n -й степени, то $f_1(x)$ — полином степени $n - 1$. Мы можем продолжить этот процесс и постепенно снизить степень результирующего полинома до нуля:

$$\frac{f_1(x) - f_1(x_0)}{x - x_0} = f_2(x), \quad (1-7.2)$$

$$\frac{f_2(x) - f_2(x_0)}{x - x_0} = f_3(x), \quad (1-7.3)$$

.....

При этом процессе мы автоматически получаем последовательные коэффициенты разложения $f(x)$ в ряд Тейлора по степеням $x - x_0$. Действительно, умножая все наши равенства на $x - x_0$ и производя последовательно подстановки, мы получим

$$f(x) = f(x_0) + f_1(x_0)(x - x_0) + f_2(x_0)(x - x_0)^2 + \dots \quad (1-7.4)$$

Коэффициенты разложения в ряд Тейлора равны, таким образом, последовательным *остаткам* ряда алгебраических делений. Этот процесс известен как «схема Горнера» [W. G. Horner (1819), а также P. Ruffini (1804)].

8. Техника подвижной полосы

При численных выкладках зачастую теряется много времени на записывание промежуточных результатов; этого можно избежать при более сжатом расположении вычислений. Один из таких приемов, очень помогающий при многих численных подсчетах, использует «подвижную полосу». Мы изложим способ подвижной полосы для случая вычислений на арифмометре, однако он может быть легко закодирован также для электронных вычислительных машин.

Некоторый фиксированный набор чисел выписывается сверху вниз на вертикальной полосе бумаги. Этот подвижной столбец используется вместе с другим заданным столбцом, с «неподвижной полосой». Операция заключается в перемножении каждой пары расположенных рядом чисел: одного числа подвижной полосы и одного — неподвижной полосы. Эти произведения складываются, и получившаяся сумма записывается в третью «нарастающую полосу», затем подвижная полоса сдвигается вертикально вниз на одну строку, операция повторяется и получается следующий элемент «нарастающей полосы». Так мы продолжаем до тех пор, пока подвижная полоса не дойдет до некоторого положения, зависящего от характера решаемой задачи, например до нижней строки неподвижной полосы. Мы получаем,

таким образом, расположение, показанное на следующей численной схеме:

Подвижная полоса	Неподвижная полоса	Нарастающая полоса	
— 3 — 2 4 1			
↓	2	2	(1-8.1)
	— 3	5	
	— 5	— 21	
	0	— 20	
	1	20	

В этом расположении порядок следования операций не имел решающего значения, ибо результаты, вписанные в «нарастающую полосу», не влияли на последующие операции. Нарастающая полоса схемы (1) могла бы быть получена при перемещении подвижной полосы снизу вверх либо в какой-нибудь другой последовательности. Часто, однако, встречается другой вид алгоритма, при котором нарастающая полоса располагается между подвижной и неподвижной полосой. При таком расположении перемножаются пары рядом расположенных чисел подвижной и нарастающей полос, и только самый нижний элемент подвижной полосы (в нашем примере 1) умножается на стоящий напротив него элемент неподвижной полосы. Операции алгоритма теперь проходят следующим образом:

Подвижная полоса	Нарастающая полоса	Неподвижная полоса	
— 3 — 2 4 1			
↓	2	2	(1-8.2)
	5	— 3	
	11	— 5	
	28	0	
	76	1	

Это — расположение «с обратной связью», при котором операции должны проводиться лишь в надлежащей последовательности.

Расположение первого рода мы встретим позже во всех расчетах, учитывающих веса заданных величин, например при локальном сглаживании, дифференцировании эмпирически заданных функций и т. д. (ср. V, 8, 10). Но при делении полиномов требуется расположение второго рода.

В качестве простого приложения техники подвижной полосы рассмотрим алгебраическое деление заданного полинома на $x - x_0$. Теперь подвижная полоса будет содержать только два элемента, а именно $x_0, 1$, расположенные по вертикали. Неподвижная полоса будет содержать последовательные коэффициенты $a_0, a_1, \dots, a_n$ заданного полинома, расположенные вертикально один под другим. Результаты последовательно выписываются между этими двумя полосами.

Мы получаем последнее число в нарастающей полосе, когда единица подвижной полосы оказывается в одной строке с коэффициентом a_0 , находящимся на неподвижной полосе. Если результат оказывается равным нулю, то $x - x_0$ есть делитель заданного полинома, и нарастающая полоса содержит последовательные коэффициенты частного $\frac{p_n(x)}{x - x_0}$. Если в результате получается число, не равное нулю, оно представляет собой остаток от деления. Это число мы не вписываем в нарастающий столбец, а переносим его в особый столбец «остатка», проставив нуль на место последнего элемента в нарастающей полосе.

Мы поясним эту технику на примере вычисления результата подстановки $x_0 = 0,79$ в кубический полином (6.1) способом алгебраического деления:

Подвижная полоса	Нарастающая полоса	Неподвижная полоса	Остаток
0,79 1			
	1,09375	1,09375	(1-8.3)
	1,9265625	1,0625	
	1,2219844	— 0,3	
	0	— 1	— 0,03463234

Схема Горнера возникает в результате последовательного повторения этого алгоритма. Каждый раз при этом «нарастающая полоса» предыдущего этапа становится «неподвижной полосой» следующего этапа. В целях более четкого расположения мы будем писать остаток всякий раз внизу соответствующей неподвижной полосы. Полная схема алгебраического деления, указанного в схеме (3), примет теперь вид

1,09375	1,09375	1,09375	1,09375
1,09375	2,790625	1,9265625	1,0625
	3,6546875	1,2219844	— 0,3
		3,4265781	— 1
			— 0,0346323

Таким образом,

$$f(0,79 + h) = 1,09375h^3 + 3,6546875h^2 + 3,4265781h - 0,0346323.$$

На практике мы часто можем остановиться после трех делений, так как $f(x_0)$, $f'(x_0)$ и $\frac{1}{2}f''(x_0)$ достаточны для применения формулы (5.12).

9. Остальные корни кубического уравнения

В нашем примере формула (5.12) дала уточненный корень (6.8). Повторяем алгебраическое деление с этим новым корнем, но не идем дальше первого этапа:

$$\begin{array}{r} 1 \\ 1,86909628 \\ 1,24892599 \\ \hline 1,041667 \\ - 0,297619 \\ \hline - 1,033399 \\ \hline - 0,00000107 \end{array} \quad (1-9.1)$$

Так как остаток уже практически пренебрежимо мал, то мы свели заданное кубическое уравнение к квадратному:

$$x^2 + 1,869096x + 1,248926 = 0. \quad (1-9.2)$$

Решая это уравнение по обычной формуле, получим

$$\left. \begin{array}{l} x_2 = -0,934548 + 0,612818i, \\ x_3 = -0,934548 - 0,612818i. \end{array} \right\} \quad (1-9.3)$$

Возвращаясь к исходному уравнению (4.1), мы вычислим обратные значения и поделим их на 0,24, после чего найдем три корня уравнения (4.1):

$$\left. \begin{array}{l} \xi_1 = 5,035677, \\ \xi_2 = -3,117839 + 2,044483i, \\ \xi_3 = -3,117839 - 2,044483i. \end{array} \right\} \quad (1-9.4)$$

10. Подстановка комплексного числа в полином

Умножение двух комплексных чисел гораздо более громоздко, чем умножение двух вещественных чисел. Поэтому подстановка комплексного числа $x_0 = a + ib$ в полином является громоздкой операцией, даже если пользоваться способом алгебраического деления. Нижеследующий метод подстановки обладает тем преимуществом, что он сводит число операций с комплексными числами к минимуму.

При обычном способе алгебраического деления мы должны разделить $f(x)$ на «корневой множитель» $x - x_0 = x - a - ib$, что сразу же приводит к необходимости перемножать комплексные числа. Разделим поэтому $f(x)$ на вещественную величину *):

$$(x - x_0)(x - x_0^*) = (x - a)^2 + b^2 = x^2 - 2ax + (a^2 + b^2). \quad (1-10.1)$$

Результат деления может быть записан следующим образом:

$$\frac{f(x)}{(x - x_0)(x - x_0^*)} = f_1(x) + \frac{Ax + B}{(x - x_0)(x - x_0^*)}. \quad (1-10.2)$$

Теперь остаток уже не постоянное число, а линейное выражение $Ax + B$. Умножив (10.2) почленно на знаменатель и положив x равным x_0 , получим

$$f(x_0) = Ax_0 + B. \quad (1-10.3)$$

Это видоизменение обычной схемы имеет, следовательно, то преимущество, что подстановка комплексного числа x_0 производится *лишь в конце* и только в выражении Ax_0 . Вплоть до этого момента все операции вещественны.

Подвижная полоса для схемы алгебраического деления теперь состоит из чисел $-(a^2 + b^2)$, $2a$, 1 . Операции проводятся по прежнему способу с той лишь разницей, что деление заканчивается за один шаг *до того*, как мы доходим до последнего коэффициента функции $f(x)$. Следовательно, процесс этот дает два остаточных коэффициента A , B . Эти коэффициенты переносим в особый столбец «остатка» справа от неподвижной полосы, а последние два элемента нарастающей полосы полагаем равными нулю.

Пример. Подставим комплексное число $x = 0,3 - 0,5i$ в

$$f(x) = x^4 - 2x^3 + 4x^2 - 2x + 1. \quad (1-10.4)$$

Подвижная полоса	Частное	Остаток	
$-0,34$ $0,6$ 1	1 $-1,4$ $2,82$	1 -2 4	(1-10.5)
↓	0 0	-2 1	
		$0,168$ $0,0412$	

) x^ обозначает комплексное число, сопряженное числу x .

Таким образом, остаток равен $0,168x + 0,0412$, и мы получаем

$$\begin{aligned} f(0,3 - 0,5i) &= 0,168(0,3 - 0,5i) + 0,0412 = \\ &= 0,0916 - 0,084i. \end{aligned} \quad (1-10.6)$$

Однако метод Горнера при этом способе *не* применим: мы должны найти $f'(x)$ и $f''(x)$ путем фактического дифференцирования и применить процесс к этим полиномам.

В нашем примере, если мы рассматриваем заданное комплексное число как предварительное значение корня x_0 , которое должно быть скорректировано по методу Ньютона, процесс продолжается следующим образом:

$$\begin{array}{r|l} \begin{array}{rr} f'(x) & \\ 4 & 4 \\ -3,6 & -6 \\ \hline 0 & 8 \\ 0 & -2 \end{array} & \begin{array}{l} 4,48 \\ -0,776 \end{array} \\ \hline \end{array} \quad \begin{array}{r|l} \begin{array}{rr} \frac{1}{2} f''(x) & \\ 6 & 6 \\ 0 & -6 \\ 0 & 4 \end{array} & \begin{array}{l} -2,4 \\ 1,96 \end{array} \\ \hline \end{array} \quad (1-10.7)$$

Поэтому

$$\begin{aligned} f'(x_0) &= 4,48(0,3 - 0,5i) - 0,776 = 0,568 - 2,24i, \\ \frac{1}{2} f''(x_0) &= -2,40(0,3 - 0,5i) + 1,96 = 1,24 + 1,2i. \end{aligned}$$

Применение формулы (5.12) дает

$$\frac{1}{h} = \frac{1,24 + 1,2i}{0,568 - 2,24i} - \frac{0,568 - 2,24i}{0,0916 - 0,084i}, \quad (1-10.8)$$

и мы получаем $h = -0,042909 - 0,029222i$.

Следовательно, уточненный корень

$$x_1 = 0,257091 - 0,529222i. \quad (1-10.9)$$

Еще раз подставляя в $f(x)$, получаем

$$\begin{array}{r|l} \begin{array}{rr} 1 & 1 \\ -1,485818 & -2 \\ 2,889847 & 4 \\ \hline 0 & -2 \\ 0 & 1 \end{array} & \begin{array}{l} 0,000256 \\ -0,000383 \end{array} \\ \hline \end{array} \quad (1-10.10)$$

Новый остаток оказывается равным

$$f(x_1) = 0,000256x_1 - 0,000383 = -0,000318 - 0,000135i.$$

Оценку точности корня x_1 можно получить следующим образом. Так как x_1 весьма близко к x_0 , то $f'(x_1)$ очень мало отличается от $f'(x_0)$.

Первая поправка по методу Ньютона требует вычисления $-\frac{f(x_1)}{f'(x_1)}$; это выражение для нужд оценки может быть заменено выражением $-\frac{f(x_1)}{f'(x_0)}$. Так как абсолютное значение $f'(x_0)$ больше 2, то остаток $f(x_1)$ показывает, что ошибка корня x_1 не может превышать 1,5 единиц четвертого десятичного знака. Если такая точность нас удовлетворяет, то мы можем считать

$$x_1 = 0,2571 - 0,5292i$$

окончательным значением корня. Тогда коэффициенты процесса деления, выполненного в (10), дают нам «приведенное уравнение»

$$x^2 - 1,4858x + 2,8898 = 0, \quad (1-10.11)$$

решение которого дает оставшуюся пару комплексных корней.

11. Уравнения четвертой степени

Алгебраические уравнения четвертой степени с корнями, вообще говоря комплексными, часто встречаются при анализе устойчивости самолетов и в проблемах, связанных с использованием сервомеханизмов. Исторически сложившийся метод решения алгебраических уравнений четвертой степени включает следующие этапы. С помощью преобразования вида $x \mapsto x + \alpha$ уничтожается коэффициент при кубическом члене. Затем решается некоторое вспомогательное кубическое уравнение. Корни первоначального уравнения строятся с помощью трех корней этого вспомогательного уравнения. При конкретных числовых расчетах этот метод оказывается очень длинным и громоздким. Нижеследующее видоизменение этого традиционного способа дает четыре корня произвольного квадратного уравнения с вещественными коэффициентами с помощью быстрого и численно удобного приема.

Любое уравнение четвертой степени

$$x^4 + c_1x^3 + c_2x^2 + c_3x + c_4 = 0 \quad (1-11.1)$$

можно представить в следующем виде:

$$(x^2 + \alpha x + \beta)^2 = (ax + b)^2. \quad (1-11.2)$$

Если первоначальные c_i вещественны, то новые коэффициенты также будут вещественными. Поэтому решение первоначального уравнения сводится к решению двух квадратных уравнений

$$x^2 + \alpha x + \beta \pm (ax + b) = 0, \quad (1-11.3)$$

которые имеют четыре (вообще говоря, комплексных) корня, получающихся по обычным формулам. Коэффициенты уравнения (3) могут

быть найдены следующим образом. Вычислим последовательно постоянные:

$$\alpha = \frac{c_1}{2}, \quad A = c_2 - \alpha^2, \quad B = c_3 - \alpha A \quad (1-11.4)$$

и составим кубическое уравнение *)

$$\xi^3 + (2A - \alpha^2)\xi^2 + (A^2 + 2B\alpha - 4c_4)\xi - B^2 = 0. \quad (1-11.5)$$

Так как левая его часть при $\xi = 0$ отрицательна, то оно должно иметь положительный корень. Мы определяем его по методу § 3. Чтобы избежать последующих поправок, целесообразно найденный приближенный корень уравнения (5) теперь же исправить по методу Ньютона (ср. § 5) и получить ξ с большой точностью. Коэффициенты приведенного уравнения (3) определяются после этого по формулам¹⁾

$$\left. \begin{aligned} \alpha &= \frac{1}{2} c_1, & \beta &= \frac{1}{2} (A + \xi), \\ a &= \sqrt{\xi}, & b &= \frac{a}{2} \left(\alpha - \frac{B}{\xi} \right). \end{aligned} \right\} \quad (1-11.6)$$

Численный пример. Мы поясним этот общий прием, решив следующее уравнение четвертой степени:

$$x^4 + 7,64x^3 + 23,6044x^2 + 38,91024x + 38,149496 = 0.$$

Здесь

$$c_1 = 7,64, \quad c_2 = 23,6044, \quad c_3 = 38,91024, \quad c_4 = 38,149496.$$

Подставляя в общие равенства (4) эти значения, получаем

$$\alpha = 3,82, \quad A = 9,012, \quad B = 4,4844.$$

Кубическое уравнение (5) принимает при этом вид

$$\xi^3 + 3,431600\xi^2 - 37,121024\xi - 20,109843 = 0.$$

Подставляя

$$\xi = \frac{\eta}{0,37},$$

*) Уравнение (5), предназначенное фактически для отыскания коэффициента a (ибо $a^2 = \xi$), выводится путем исключений величин α, β, b из системы четырех уравнений, получаемых приравниванием коэффициентов при степенях x в уравнениях (1) и (2).

¹⁾ Случай $B = 0$ требует особого рассмотрения. Уравнение (5) имеет тогда решение $\xi = 0$, и с помощью перехода к пределу мы получаем

$$\alpha = \frac{1}{2} c_1, \quad \beta = \frac{1}{2} A, \quad a = 0, \quad b = \frac{1}{2} \sqrt{A^2 - 4c_4}.$$

Если, однако, $A^2 - 4c_4$ при этом оказывается отрицательным, то следует разделить (5) на ξ и найти положительный корень полученного квадратного уравнения. Это значение ξ подставляется тогда в общие формулы (6) при $B = 0$.

получаем

$$\eta^3 + 1,269692\eta^2 - 5,081868\eta - 1,018624 = 0.$$

Так как левая часть этого уравнения также отрицательна при $\eta = 1$, то мы обращаем его, переходя к обратному корню (попутно делим уравнение на 1,018624):

$$\begin{array}{r} \bar{\eta}^3 + 4,988954\bar{\eta}^2 - 1,246478\bar{\eta} - 0,981717 = 0 \\ \swarrow \\ 1,5 \quad \quad - 0,5625 \quad + 0,03125 \\ \hline 6,488954\bar{\eta}_0^2 - 1,808978\bar{\eta}_0 - 0,950467 = 0, \\ \bar{\eta}_0^2 - 0,2788\bar{\eta}_0 - 0,1465 = 0, \\ \bar{\eta}_0 = 0,1394 + \sqrt{0,1658} = 0,5466. \end{array}$$

Этот корень мы исправим при помощи метода Ньютона, пользуясь алгебраическим делением по схеме Горнера (ср. §§ 7 и 8).

Покажем эту схему в горизонтальном расположении:

1	4,988954	— 1,246478	— 0,981717	
1	5,535554	1,779256	— 0,009176	
1	6,082154	5,103761		
1	6,628754			
$\frac{1}{h}$	$= \frac{5,103761}{0,009176}$	$+ \frac{6,628754}{5,103761}$	$=$	$\frac{556,2076}{+ 1,2988}$
				$\frac{557,5064}{h = 0,001794}$
				$\frac{0,5466}{\bar{\eta} = 0,548394}$

$$\bar{\eta} = 0,548394, \quad \eta = 1,823506, \quad \xi = 4,928394.$$

Подстановка в равенства (6) дает

$$\alpha = 3,82, \quad \beta = 6,970197, \quad a = 2,219998, \quad b = 3,230196.$$

Следовательно, приведенное квадратное уравнение в данном случае имеет вид

$$\begin{aligned} & x^2 + 3,82x + 6,970197 \pm (2,219998x + 3,230196) = 0; \\ (a) \quad & x^2 + 6,039998x + 10,200393 = 0, \\ & x = -3,019999 \pm 1,039230i \text{ (точно: } -3,02 \pm 1,03923048i); \\ (b) \quad & x^2 + 1,600002x + 3,740001 = 0, \\ & x = -0,800001 \pm 1,760681i \text{ (точно: } -0,8 \pm 1,76068169i). \end{aligned}$$

Однако достаточно бóльшую точность можно получить даже при условии, что мы не будем исправлять предварительное значение $\bar{\eta}_0$, а примем его в качестве $\bar{\eta}_1$. Тогда

$$\bar{\eta}_1 = 0,5466, \quad \eta_1 = 1,8295, \quad \xi = 4,9446,$$

и подстановка в (6) приводит теперь к значениям

$$\alpha = 3,82, \quad \beta = 6,9783, \quad a = 2,2236, \quad b = 3,2388,$$

которые дают приведенное уравнение

$$x^2 + 3,82x + 6,9783 \pm (2,2236x + 3,2388) = 0.$$

Четыре корня этого уравнения суть

$$x = -3,0218 \pm 1,0420i \quad \text{и} \quad x = -0,7982 \pm 1,7614i.$$

Хороший контроль точности, на которую можно при этом рассчитывать, мы получим, образовав произведение всех четырех корней. Оно должно равняться c_4 . В нашем случае произведение первых четырех корней дает 38,149461 (в то время как $c_4 = 38,149496$). Вторая группа корней дает 38,2081.

12. Уравнения высших степеней

Метод поправок Ньютона является важным орудием для постепенного уточнения корня алгебраического уравнения, если мы можем начать с какого-нибудь подходящего первоначального приближения. К сожалению, мы не обладаем каким-либо прямым методом для приближенной локализации корней уравнений выше четвертой степени. *Вещественные* корни алгебраического уравнения могут быть найдены с затратой относительно небольшого труда. С этой целью мы делим интервал между -1 и $+1$, скажем, на 10 равных частей и вычисляем значения $f(x)$ для точек, отстоящих друг от друга на 0,2. Мы замечаем изменение знака и локализуем более точно корень путем линейной интерполяции. Схема алгебраического деления, описанная в § 7, уточнит затем этот корень с желаемой степенью точности. Корни, лежащие вне интервала $(-1, +1)$, следует ввести внутрь интервала путем обратного преобразования (3.7), т. е. обращая порядок следования коэффициентов. Затем следует вновь изучить этот интервал на каждом его участке длиной в 0,2. Таким образом, сравнительно несложные расчеты приводят к локализации всех вещественных корней полинома.

Однако корни полинома, вообще говоря, не обязательно будут вещественны, а полиномы, встречающиеся в задачах колебаний и

устойчивости, обычно имеют комплексные корни. Для уравнений такого типа большое значение имеет так называемый «метод моментов», впервые описанный Даниилом Бернулли (1728). Этот метод дает возможность сравнительно легко локализовать наибольший по модулю корень алгебраического уравнения любой степени.

13. Метод моментов

Если мы разложим полином n -й степени на линейные множители

$$f(x) = a_0 x^n + a_1 x^{n-1} + \dots + a_n \equiv \equiv a_0 (x - x_1)(x - x_2) \dots (x - x_n) \quad (1-13.1)$$

и возьмем его логарифмическую производную, то получим формулу, особенно полезную для изучения алгебраического поведения полинома:

$$-\frac{f'(x)}{f(x)} = \frac{1}{x_1 - x} + \frac{1}{x_2 - x} + \dots + \frac{1}{x_n - x}. \quad (1-13.2)$$

Пусть $x = x_0$ есть точка, расположенная недалеко от корня $x = x_1$. Тогда

$$-\frac{f'(x_0)}{f(x_0)} = \frac{1}{x_1 - x_0} + \frac{1}{x_2 - x_0} + \dots + \frac{1}{x_n - x_0}, \quad (1-13.3)$$

и мы видим, что метод Ньютона вычисления поправки для x_0 по формуле

$$\frac{1}{h} = -\frac{f'(x_0)}{f(x_0)} \quad (1-13.4)$$

сводится к тому, чтобы сохранить в правой части (3) только первый ее член. Отбрасывание остальных членов будет тем более оправданным, чем ближе x_0 к истинному корню x_1 .

В частности, при $x_0 = 0$ из (3) получаем

$$-\frac{f'(0)}{f(0)} = \frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}. \quad (1-13.5)$$

Если случится, что один корень, скажем x_1 , намного ближе к нулю, чем все остальные корни, то

$$\frac{1}{h} = -\frac{f'(0)}{f(0)} \quad (1-13.6)$$

даст хорошее приближение к этому корню. Однако ценность этого приближения теряется все более по мере того, как близость x_1 к нулю становится все менее четко выраженной.

Продифференцируем теперь функцию (2) $m - 1$ раз. Мы получим тогда

$$-\frac{1}{(m-1)!} \left[\frac{f'(x)}{f(x)} \right]^{(m-1)} = \frac{1}{(x_1 - x)^m} + \dots + \frac{1}{(x_n - x)^m}, \quad (1-13.7)$$

а положив $x = 0$, получим:

$$-\frac{1}{(m-1)!} \left[\frac{f'(x)}{f(x)} \right]_{x=0}^{(m-1)} = \frac{1}{x_1^m} + \dots + \frac{1}{x_n^m}. \quad (1-13.8)$$

Таким путем ближайший к нулю корень x_1 выделяется все более четко, даже если близость x_1 к нулю была первоначально не очень отчетливо выражена.

Мы приходим, таким образом, к следующему обобщению формулы (6):

$$\frac{1}{h^m} = -\frac{1}{(m-1)!} \left[\frac{f'(x)}{f(x)} \right]_{x=0}^{(m-1)}. \quad (1-13.9)$$

Выбирая m достаточно большим, отсюда можно получить хорошее приближение к наименьшему по модулю корню x_1 .

Отправляясь от этой общей идеи, можно разработать ценный метод локализации наименьшего по модулю корня алгебраического уравнения. Однако фактическое дифференцирование функции (2) было бы делом весьма кропотливым. Можно все же получить выражения (9) с помощью простой схемы, обобщающей метод алгебраического деления, рассмотренного в §§ 8 и 9.

Предпочтительнее при этом иметь дело с корнями *обращенного* полинома, которые обратны первоначальным корням.

Тогда нахождение ближайшего к нулю корня заменяется нахождением корня, наиболее удаленного от нуля. Правая часть формулы (8) принимает в этом случае вид

$$S_m = x_1^m + x_2^m + \dots + x_n^m. \quad (1-13.10)$$

Величины S_m называют «симметрическими моментами корней» или «степенными суммами».

Формула (9) записывается теперь в виде

$$\frac{1}{h^m} = x_1^m + \dots + x_n^m = S_m. \quad (1-13.11)$$

В следующем параграфе будет дана простая и изящная численная схема для последовательного вычисления моментов S_m .

14. Алгебраическое деление двух полиномов

Способ подвижной полосы может быть с успехом применен для вычисления симметрических функций S_m от корней полинома *).

Выпишем коэффициенты полинома $A(x)$ в вертикальный столбец, начиная со старшего коэффициента и кончая младшим («неподвижная

*) Симметрическими функциями переменных $x_1, x_2, \dots, x_n$ называются такие функции, которые остаются неизменными при любой перестановке этих переменных.

полоса»). Пусть, далее, $B(x)$ — полином, у которого коэффициент старшего члена нормирован к 1. Мы выписываем коэффициенты $B(x)$ на подвижную полосу, начиная с коэффициента, нормированного к 1, располагая их снизу вверх. При этом знак каждого коэффициента меняем на обратный, за исключением старшего коэффициента 1.

Способом подвижной полосы образуем теперь частное от деления полиномов $A(x)$ и $B(x)$.

Пример. Разделить $6x^6 - 3x^5 + x^3 - x^2$ на $x^3 + 2x^2 + 4x - 1$.

$B(x)$	Частное	$A(x)$	Остаток
1 — 4 — 2 1	6	6	
	— 15	— 3	
	+ 6	0	
	55	1	(1-14.1)
	0	— 1	— 150
	0	0	— 214
	0	0	55

Следовательно,

$$\frac{6x^6 - 3x^5 + x^3 - x^2}{x^3 + 2x^2 + 4x - 1} = 6x^3 - 15x^2 + 6x + 55 + \frac{-150x^2 - 214x + 55}{x^3 + 2x^2 + 4x - 1}.$$

Применение этого приема к нахождению дроби $\frac{f'(x)}{f(x)}$ наталкивается на то затруднение, что числитель имеет степень $n - 1$, а знаменатель — степень n . Это затруднение можно обойти, умножая числитель на произвольно высокую степень x^N . Тогда процесс «деления» дает в результате полином степени $N - 1$. Если найденный полином мы разделим на x^N , то получим частное в форме

$$\frac{f'(x)}{f(x)} = \frac{c_0}{x} + \frac{c_1}{x^2} + \frac{c_2}{x^3} + \dots, \tag{1-14.2}$$

где коэффициенты $c_0, c_1, c_2, \dots$ суть последовательные величины из столбца «частного».

С другой стороны, разложение ряда (13.2) по обратным степеням x дает

$$\begin{aligned} \frac{f'(x)}{f(x)} &= \frac{1}{x} \left(\frac{1}{1 - \frac{x_1}{x}} + \dots + \frac{1}{1 - \frac{x_n}{x}} \right) = \\ &= \frac{n}{x} + \frac{S_1}{x^2} + \frac{S_2}{x^3} + \dots + \frac{S_m}{x^{m+1}}. \end{aligned} \tag{1-14.3}$$

Сравнение рядов (2) и (3) показывает, что

$$c_m = S_m. \tag{1-14.4}$$

Таким образом, метод подвижной полосы дает нам последовательные степенные суммы вплоть до любого порядка m .

Пример. Найти последовательные степенные суммы полинома

$$x^4 + 7,60x^3 + 23,34x^2 + 38,44x + 37,40. \quad (1-14.5)$$

$f(x)$			
m	$\begin{array}{r} -37,40 \\ -38,44 \\ -23,34 \\ -7,60 \\ 1 \end{array}$	S_m	$f'(x)$
	0	4	
	1	$-7,60$	4
	2	11,08	22,80
	3	$-22,144$	46,68
4	↓	52,2312	38,44
5		$-616,6382$	
6		...	

(1-14.6)

15. Степенные суммы и наибольший по модулю корень

Характер изменения моментов высших порядков позволяет сделать определенное заключение о модуле наиболее далекого от нуля корня.

Запишем корни, вообще говоря комплексные, в полярной форме:

$$x_k = r_k e^{i\theta_k}. \quad (1-15.1)$$

Тогда

$$S_m = r_1^m e^{im\theta_1} \left[1 + \left(\frac{r_2}{r_1} \right)^m e^{im(\theta_2 - \theta_1)} + \dots \right. \\ \left. \dots + \left(\frac{r_n}{r_1} \right)^m e^{im(\theta_n - \theta_1)} \right]. \quad (1-15.2)$$

Если r_1 — модуль наиболее удаленного корня, а все остальные r_i меньше r_1 , то дроби $\frac{r_i}{r_1}$ ($i = 2, \dots, n$) при возведении в возрастающие степени стремятся к нулю, и поэтому для очень больших m имеем приближенно:

$$S_m = r_1^m e^{im\theta_1}$$

и

$$r_1 = \sqrt[m]{|S_m|}. \quad (1-15.3)$$

Таким образом, последовательные степенные суммы обладают тем ценным свойством, что они выделяют наибольший по модулю корень с возрастающей силой.

Если корни алгебраического уравнения мало отличаются по модулю, то это различие значительно возрастает при переходе к суммам достаточно высоких степеней. Например, отношение $1,5:1$ после десяти этапов возрастает до отношения $1,5^{10}:1 = 57,7$. Наибольший по модулю корень будет, следовательно, «заглушать» все остальные корни. Возможно, однако, что несколько корней будут иметь один и тот же модуль r_1 , больший чем модули остальных корней. Действительно, если алгебраическое уравнение имеет вещественные коэффициенты, то мы заранее знаем, что комплексные корни появляются парами $a + ib$ и $a - ib$ и, следовательно, по меньшей мере, два корня лежат на окружности максимального радиуса $r = r_1$. Тогда последовательные степенные суммы S_m выделяют одну *пару* корней.

Вообще, величина S_m при большом m в основном будет определяться лишь наибольшими по модулю корнями, тогда как меньшие по модулю корни практически роли играть не будут.

Если наибольший по модулю корень вещественный, то его можно быстро получить с помощью отношения двух последовательных сумм:

$$x = \frac{S_{m+1}}{S_m}. \quad (1-15.4)$$

Простое рассмотрение нескольких последовательных S_m помогает выявить в этом случае такое положение. Суммы S_m либо все положительны, либо поочередно меняют знаки. Кроме того, отношение (4) остается приблизительно одинаковым, если брать два соседних отношения $\frac{S_{m+1}}{S_m}$ и $\frac{S_{m+2}}{S_{m+1}}$.

В более часто встречающемся случае комплексных корней такая регулярность в величине и знаке сумм S_m не может наблюдаться. За большим значением может следовать малое, а знаки могут причудливо чередоваться. Такого рода поведение указывает, что наибольший по модулю корень — комплексный: $a \pm ib$. Соответствующее S_m теперь будет иметь вид

$$S_m = 2r^m \cos m\theta. \quad (1-15.5)$$

Множитель $\cos m\theta$ ответствен за неправильное чередование знаков и модулей величин S_m . Однако, если на окружности максимального радиуса находится только одна пара комплексных корней, то мы все же можем добиться приближенной локализации наибольшего по модулю корня.

Рассмотрим уравнение, заданное в форме определителя:

$$\begin{vmatrix} 1 & \lambda & \lambda^2 & \dots & \lambda^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{vmatrix} = 0, \quad (1-15.6)$$

которое удовлетворяется при $\lambda = x_1, x_2, \dots, x_n$. Если этот определитель умножить слева на определитель

$$\begin{vmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & w_1 & w_2 & \dots & w_n \\ 0 & w_1 x_1 & w_2 x_2 & \dots & w_n x_n \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & w_1 x_1^{n-1} & w_2 x_2^{n-1} & \dots & w_n x_n^{n-1} \end{vmatrix}, \quad (1-15.7)$$

то мы получаем уравнение

$$\begin{vmatrix} 1 & \lambda & \lambda^2 & \dots & \lambda^n \\ w_0 & w_1 & w_2 & \dots & w_n \\ w_1 & w_2 & w_3 & \dots & w_{n+1} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ w_{n-1} & w_n & w_{n+1} & \dots & w_{2n-1} \end{vmatrix} = 0, \quad (1-15.8)$$

где w_k обозначают «взвешенные степенные суммы»

$$w_k = w_1 x_1^k + w_2 x_2^k + \dots + w_n x_n^k. \quad (1-15.9)$$

Будем теперь составлять уравнения типа (8) последовательными этапами, начиная с квадратного уравнения

$$\begin{vmatrix} 1 & \lambda & \lambda^2 \\ w_0 & w_1 & w_2 \\ w_1 & w_2 & w_3 \end{vmatrix} = 0 \quad (1-15.10)$$

и добавляя каждый раз по две строки и по два столбца определителя (8). При этой последовательной процедуре каждый новый шаг порождает дополнительную пару корней и в то же время корректирует ранее полученные корни. Вообще говоря, мы не можем ожидать, что эти поправки будут малыми. Но это действительно будет иметь место, если веса $w_1, w_2, \dots$ особенно сильно выделяют наибольшие по модулю корни в выражениях w_k . Если веса w_1, w_2 превалируют по сравнению с остальными весами, то решение простого квадратного уравнения (10) уже достаточно хорошо аппроксимирует наибольшую по модулю пару корней, и дальнейшие шаги дадут лишь небольшие поправки.

Допустим теперь, что наибольший по модулю корень превосходит другие на достаточно большой множитель, например 1,5 или более. Рассмотрим четыре последовательные степенные суммы S_m , соответствующие индексам $m, m+1, m+2, m+3$, где m не

меньше 9. Эти четыре суммы S_m можно рассматривать в качестве четырех ω_i уравнения (10):

$$\left. \begin{aligned} \omega_0 &= \omega_1 + \omega_2 + \dots + \omega_n, \\ \omega_1 &= \omega_1 x_1 + \omega_2 x_2 + \dots + \omega_n x_n, \\ \omega_2 &= \omega_1 x_1^2 + \omega_2 x_2^2 + \dots + \omega_n x_n^2, \\ \omega_3 &= \omega_1 x_1^3 + \omega_2 x_2^3 + \dots + \omega_n x_n^3, \end{aligned} \right\} \quad (1-15.11)$$

где

$$\omega_1 = x_1^m, \quad \omega_2 = x_2^m, \quad \dots, \quad \omega_n = x_n^m. \quad (1-15.12)$$

Условия применимости простого квадратного уравнения (10), таким образом, выполняются. Образует отношения

$$s_1 = \frac{S_m}{S_{m+1}}, \quad s_2 = \frac{S_{m+2}}{S_{m+1}}, \quad s_3 = \frac{S_{m+3}}{S_{m+1}} \quad (1-15.13)$$

и решаем квадратное уравнение

$$\begin{vmatrix} 1 & \lambda & \lambda^2 \\ s_1 & 1 & s_2 \\ 1 & s_2 & s_3 \end{vmatrix} = 0, \quad (1-15.14)$$

или

$$(1 - s_1 s_2) \lambda^2 + (s_1 s_3 - s_2) \lambda + (s_2^2 - s_3) = 0. \quad (1-15.15)$$

Корни (обычно комплексные) этого уравнения определяют наибольшие по модулю корни с достаточной точностью. В случае вещественных корней мы берем лишь наибольший из двух корней.

Поясним численно этот метод, применив его к примеру из § 14. Если продолжить вычисление с подвижной полосой таблицы (1-14.6) до $m = 12$, то последние четыре степенные суммы будут:

$$\begin{aligned} S_9 &= 61829,6, & s_1 &= -0,31103, \\ S_{10} &= -198790,6, & \text{отсюда } s_2 &= -2,91902, \\ S_{11} &= 580274,3, & s_3 &= 7,55682. \\ S_{12} &= -1502225, \end{aligned}$$

Квадратное уравнение (15) в этом случае имеет вид

$$\begin{aligned} 0,092097\lambda^2 + 0,56862\lambda + 0,96386 &= 0, \\ \lambda &= -3,087 \pm 0,967i. \end{aligned}$$

Точный корень уравнения (14.5)

$$x = -3 \pm i,$$

и ошибка, как видим, не превышает 3% . Мы получили, следовательно, корень, который настолько близок к истинному, что для его уточнения можно применить метод Ньютона (ср. § 5).

16. Оценка наибольшего абсолютного значения

Предположение, что абсолютное значение наибольшего по модулю корня будет по меньшей мере в 1,5 раза больше ближайшего корня, не всегда верно. Надо быть готовым к тому, что две или даже больше пар комплексных корней будут лежать почти на одной окружности комплексной плоскости. В особом случае все комплексные корни алгебраического уравнения могут иметь почти одинаковое абсолютное значение. Тем не менее даже в таких случаях степенные суммы S_m содержат ценную информацию относительно расположения наибольших по модулю корней. Радиус максимальной окружности можно определить с достаточной точностью, даже если две и более пар корней лежат близ этой окружности. Соответственно этому мы будем действовать следующим образом. Мы получаем сначала близкое значение наибольшего модуля r пары или нескольких пар комплексных корней. Затем приступаем к локализации угла θ или нескольких углов θ_i , соответствующих комплексным корням, лежащим близ максимальной окружности. Мы рассмотрим два способа решения этой второй части проблемы. Один основан на свойствах «скрытых периодичностей», тогда как второй использует так называемое «преобразование обратными радиусами».

Множитель $\cos m\theta$ в выражении (15.5) затрудняет простое определение радиуса r с помощью степенных сумм S_m . При отсутствии этого множителя мы могли бы получить r с помощью равенства

$$\log r = \frac{1}{m} \log \frac{|S_m|^*}{2}. \quad (1-16.1)$$

Однако, приняв некоторые меры предосторожности, мы сможем все же воспользоваться этим равенством, несмотря на мешающий множитель. Каково бы ни было θ , неизбежно случится, что $m\theta$ окажется близким к некоторому кратному величины π . В этом случае множитель $\cos m\theta$ будет близким к ± 1 . Мы можем уловить эти удобные случаи, отмечая «пиковые» значения сумм S_m . Начиная с $m=10$, мы будем вычислять логарифмы величин $|S_m|$ и делить их на m . Дойдем до $m=16$ и выберем наибольшее из этих частных. Для этого максимума образуем выражение (1) и получим r . Этот метод применим даже в случае, когда больше чем одна пара комплексных корней оказывается вблизи максимальной окружности. Даже

*) Точнее, следует взять равенство

$$\log r = \frac{1}{m} \log \frac{|S_m|}{k},$$

где k — число корней, находящихся на максимальной окружности, однако, если k неизвестно, то можно пользоваться и формулой (1), так как при $m \rightarrow \infty$ оба выражения стремятся к одному пределу.

ошибка в 100% при определении S_{16} , вызванная, например, наличием второй пары комплексных корней, исказит значение r не больше чем на 4%, так как корень шестнадцатой степени из 2 есть 1,0443. Следовательно, всегда можно будет получить достаточно точное значение для радиуса максимальной окружности, не составляя таблицу для S_m до чрезмерно больших степеней.

Численный пример поможет выяснить общий ход этого метода. Возьмем уравнение шестой степени, корни которого известны. Уравнение

$$\xi^6 - 2\xi^5 + 6\xi^4 - 58\xi^3 + 633\xi^2 - 200\xi + 2500 = 0 \quad (1-16.2)$$

имеет следующие три пары сопряженных комплексных корней:

$$\xi = \pm 2i, \quad -3 \pm 4i, \quad 4 \pm 3i.$$

В соответствии с нашими общими правилами мы нормируем свободный член к величине, близкой к единице. С этой целью мы делим коэффициенты уравнения на последовательные степени 4, выполняя подстановку

$$\xi = 4x. \quad (1-16.3)$$

Это дает новое уравнение

$$x^6 - 0,5x^5 + 0,375x^4 - 0,90625x^3 + 2,47266x^2 - 0,19351x + 0,61035 = 0. \quad (1-16.4)$$

Методом подвижной полосы, описанным в § 14, получаем последовательно следующие степенные суммы:

$$\begin{array}{cccc} 6, & 0,5, & -0,5, & 2,28125, & -8,10939, & -5,623069, & . \\ -0,0312454, & -11,299692, & 10,068455, & 20,171028. & (1-16.5) \\ -0,001920, & \underline{32,925676}, & 7,659814, & \end{array}$$

Пиковое значение $S_{11} = 32,925676$ приводит к особенно большому r ; поэтому мы на нем останавливаемся; остальные S_m опускаем. Делим на 2 и берем логарифм

$$\log r_0 = \frac{1}{11} \log 16,463 = 0,1106.$$

Это дает значение

$$r_0 = 1,290,$$

которое является удовлетворительным приближением к точному $r = 1,25$.

17. Развертка единичной окружности *)

Комплексное число

$$z = re^{i\theta}$$

характеризуется модулем и направлением. Даже если мы нашли r , нам еще необходимо найти угол θ , соответствующий комплексному корню. Если вблизи максимальной окружности расположено несколько корней, то надо найти все соответствующие θ_i . Кроме того, мы стараемся улучшить предварительное значение r , найденное только что указанным методом пиковых значений.

Нашим первым шагом будет перевод максимальной окружности в единичную окружность путем масштабного преобразования

$$x = r_0 u. \quad (1-17.1)$$

Очевидно, моменты, отнесенные к новой переменной u , равняются прежним S_m , деленным на r_0^m . В численном примере мы нашли для r_0 значение 1,29, которое можно округлить до 1,3. Поэтому мы делим значения S_m таблицы (16.5) на последовательные степени 1,3.

Останавливаясь на S_{12} , получаем, таким образом, следующую новую таблицу:

$$\begin{array}{cccccc} 6, & 0,38462, & -0,29586, & 1,03835, & -2,83932, & \\ -1,51445, & 0,00647, & -1,8008, & 1,2343, & 1,90212. & (1-17.2) \\ -0,00014, & 1,83720, & 0,32877, & & & \end{array}$$

Если принять теперь, что радиус новой максимальной окружности равен в точности 1 (что в действительности верно лишь приближенно), то последовательные моменты, соответствующие максимальному корню, будут

$$2, \quad 2 \cos \theta, \quad 2 \cos 2\theta, \quad 2 \cos 3\theta, \quad \dots, \quad 2 \cos m\theta.$$

Если на единичной окружности лежит несколько корней, то каждому из них соответствует такой ряд, и сумма S_m принимает вид

$$S_m = 2 \cos m\theta_1 + 2 \cos m\theta_2 + \dots + 2 \cos m\theta_p. \quad (1-17.3)$$

На практике p редко превышает 2 или 3, ибо степень уравнения редко превышает 6. Однако в нашем рассуждении мы можем оставить p произвольным.

Будем теперь рассматривать величины S_m как значения функции $S(t)$ от непрерывной переменной t в определенных точках $t = m$. Эту функцию мы определяем следующим образом:

$$S(t) = 2 \cos \theta_1 t + 2 \cos \theta_2 t + \dots + 2 \cos \theta_p t. \quad (1-17.4)$$

*) Чтение этого параграфа потребует от читателя предварительного ознакомления с материалом, изложенным в § 13 гл. IV.

Отметим, что $S(t)$ составлена из чисто *периодических* слагаемых. Следовательно, задача определения корней алгебраического уравнения сводится к разысканию «скрытых периодичностей» функции, заданной в равноотстоящих точках. Мы позже займемся этой проблемой подробно (ср. IV, 22). В общей проблеме каждая периодическая компонента имеет свою собственную амплитуду, фазу и частоту. В нашем случае амплитуда фиксирована заранее числом 2; кроме того, «фазовый» аспект не имеет отношения к проблеме, так как в наше рассмотрение входят только косинусы. Частоты ω_i интересуют нас прежде всего. Различные ω_i периодических компонент соответствуют неизвестным углам $\theta_1, \theta_2, \dots, \theta_r$ вычисляемых корней.

Мы воспользуемся здесь результатами исследований, которые будут изложены в гл. IV, применяя их к рассматриваемой проблеме. Позже будет показано, насколько благоприятно гасит « σ -сглаживание» усложняющие «колебания Гиббса». Применение этих σ -множителей требует дополнительного труда умножения каждой суммы S_m на вычисленный заранее множитель. Однако фокусирующее влияние метода настолько возрастает благодаря этой сглаживающей операции, уменьшающей взаимную интерференцию различных корней на единичной окружности, что эта добавочная работа хорошо себя оправдывает.

При вычислении моментов мы идем не дальше S_{12} . Кроме того, весовой множитель для S_{12} равен 0. Поэтому приходится произвести лишь 11 умножений. Множители σ представляются следующей таблицей:

m	σ_m	m	σ_m
0	1	7	0,52708
1	0,98862	8	0,41350
2	0,95493	9	0,30010
3	0,90032	10	0,19099
4	0,82699	11	0,08987
5	0,73791	12	0
6	0,63662		

(1-17.5)

Таким образом, таблица (2) значений S_m еще раз видоизменяется. Мы умножаем S_m последовательно на соответствующие множители σ_m таблицы (5). Результаты умножений S'_m располагаем в следующем порядке. Начинаем с $S_0=6$, но делим на 2 и записываем $S'_0=3$. Затем в той же горизонтальной строке записываем $S'_1, S'_2, \dots$, пока не доходим до S'_7 . Затем мы продолжаем записи на второй строке, но теперь записываем произведения в *обратном порядке*. Таким

образом, S'_7 станет в одном столбце с S'_5 , S'_8 — с S'_4 , наконец, $\frac{1}{2} S'_{12} = 0$ — с $\frac{1}{2} S'_0 = 3$. Затем составляем суммы и разности этих двух строк. В нашем численном примере схема записи выглядит так:

$$\begin{array}{r}
 3, 0,38024, -0,28252, 0,93484, -2,34812, -1,11650, -0,00412 \mid \\
 0, 0,16511, -0,00003, 0,57084, 0,51037, -0,94828, \quad \leftarrow \mid \\
 \hline
 \text{Сумма} \quad 3, 0,54535, -0,28255, 1,50568, -1,83775, -2,06478, -0,00412 \\
 \text{Разность} \quad 3, 0,21453, -0,28249, 0,36400, -2,85849, -0,16822,
 \end{array} \quad (1-17.6)$$

Так как полуокружность разделена на 12 частей, то мы сейчас изучим единичную окружность в интервалах по 15° . Сначала, однако, мы проведем исследование в интервалах по 30° . Умножим строку «суммы» на матрицу косинусов (ср. (4-13.3)), которая фактически совпадает с матрицей A_6 из IV, 13. Эта матрица имеет заранее вычисленную таблицу элементов, и мы получаем следующую схему:

k	3	0,54535	-0,28255	1,50568	-1,83775	-2,06478	-0,00412
0	1	1	1	1	1	1	1
2	1	0,866	0,5	0,	-0,5	-0,866	-1
4	1	0,5	-0,5	-1,	-0,5	0,5	1
6	1	0	-1	0	1	0	-1
8	1	-0,5	-0,5	1	-0,5	-0,5	1
10	1	-0,866	0,5	0	-0,5	-0,866	-1
12	1	-1	1	-1	1	-1	1
0,86183 <u>6,04209</u> 1,79063 1,44892 <u>6,32142</u> 1,64511 0,88933							

Последняя строка содержит результирующие произведения верхнего ряда на последовательные строки матрицы *). Сразу же мы замечаем ясно выраженные максимумы, соответствующие $k=2=60^\circ$ и $k=8=150^\circ$. Эти максимумы указывают на существование двух комплексных корней близ единичной окружности.

Теперь мы уточним наше исследование, сократив основной интервал до 15° . При этом в расчеты вводится «разностный» ряд таблицы (6). Прежняя матрица множителей заменяется новой предварительно вычисленной матрицей. Нет надобности приводить полную схему умножения, а следует произвести только те умножения, которые дадут нам по два произведения, смежных с ранее найденными максимумами; следовательно, в нашей задаче нам нужны будут два

*) Точнее: первый, второй и т. д. элементы последнего ряда представляют собой сумму произведений элементов верхнего ряда на соответствующие элементы первой, второй и т. д. строк матрицы.

произведения для $k = 1, 3$ и аналогично два произведения для $k = 7, 9$. Численная схема имеет следующий вид:

k	3	0,21453	-0,28249	0,36400	-2,85849	-0,16822
1	1	0,9659	0,866	0,7071	0,5	0,2588
3	1	0,7071	0	-0,7071	-1	-0,7071
5	1	0,2588	-0,866	-0,7071	0,5	0,9659
7	1	-0,2588	-0,866	0,7071	0,5	-0,9659
9	1	-0,7071	0	0,7071	-1	0,7071
11	1	-0,9659	0,866	-0,7071	0,5	-0,2588
		1,74718	5,87175	2,17974	5,84523	

Оба максимума рассчитываются совершенно независимо. Сначала займемся максимумом при $k = 2$ и его двумя соседями. Три последовательные ординаты суть

$$y_0 = 1,74718, \quad y_1 = 6,04209, \quad y_2 = 5,87175.$$

Для того чтобы проинтерполировать положение точного максимума, проводим через эти три ординаты параболу второй степени и определяем точку максимума. Решение этой простой алгебраической задачи дает следующие результаты. Абсцисса максимума равна

$$x_m = \frac{1}{2} \frac{y_2 - y_0}{2y_1 - (y_0 + y_2)}, \quad (1-17.7)$$

а соответствующая ей ордината

$$y_m = y_1 + \frac{1}{4} x_m (y_2 - y_0). \quad (1-17.8)$$

В нашем численном примере

$$x_m = \frac{1}{2} \frac{4,12457}{4,46525} = 0,46185,$$

$$y_m = 6,04209 + \frac{1}{4} 0,46185 \cdot 4,12457 = 6,51832.$$

Для ординат трех последовательных заданных точек в формуле (7) приняты индексы — 1, 0, 1. В действительности же расстояние между двумя соседними абсциссами составляет 15° и средняя ордината соответствует углу $k = 2 = 30^\circ$. Следовательно, угловое значение x_m составляет

$$\theta_m = 30^\circ + 0,46185 \cdot 15^\circ = 36^\circ,928.$$

Такой же расчет, проведенный для второго максимума при $k = 8$, дает следующие результаты:

$$y_0 = 2,17974, \quad y_1 = 6,32142, \quad y_2 = 5,84523,$$

$$x_m = \frac{1}{2} \frac{3,66549}{4,61787} = 0,39688, \quad \theta_m = 120^\circ + 0,39688 \cdot 15^\circ = 125^\circ,95,$$

$$y_m = 6,32142 + \frac{1}{4} 0,39688 \cdot 3,66549 = 6,68511.$$

Максимальная ордината y_m может быть использована для более точного определения модуля r корня. Если бы предположение $r=1$ было верно, то y_m должно было бы равняться

$$\frac{1}{2} + \sigma_1 + \sigma_2 + \dots + \sigma_{11} = 7,0669.$$

Но для произвольного значения r мы получаем более общую формулу

$$y_m = \frac{1}{2} + \sigma_1 r + \sigma_2 r^2 + \dots + \sigma_{11} r^{11} = A(r). \quad (1-17.9)$$

Нетрудно составить численную таблицу функции $A(r)$, в которой значения r следуют через достаточно малые интервалы в пределах r от 0,8 до 1,2. Тогда надлежащее значение r можно получить линейной интерполяцией. Эта таблица дана в приложении (табл. I). С помощью табличных значений находим, что y_m первого максимума соответствует

$$r = 0,9802,$$

тогда как y_m второго максимума соответствует

$$r = 0,9864.$$

Возвращаемся назад к первоначальному переменному x , для чего, согласно (1), требуется умножение на $r_0 = 1,3$. Получаем, таким образом, первую пару комплексных корней уравнения (15.4) в виде

$$x = 0,9802 \cdot 1,3 (\cos 36^\circ,93 \pm i \sin 36^\circ,93) = 1,018 \pm 0,765i.$$

Эти значения x хорошо согласуются с точными значениями корней $1 \pm 0,75i$ (ошибка около $1,5\%$).

Вторая пара комплексных корней получается в виде

$$x = 0,9864 \cdot 1,3 (\cos 125^\circ,95 \pm i \sin 125^\circ,95) = -0,753 \pm 1,038i,$$

тогда как точное значение равно $-0,75 \pm i$; здесь ошибка составляет 3% .

18. Преобразование обратными радиусами

Теперь мы рассмотрим второй, численно даже более быстрый прием локализации корней вдоль окружности наибольшего радиуса. Как и раньше, мы начинаем с оценки радиуса максимальной окружности методом пиковых значений § 16. Затем мы выполняем преобразование (17.1). Теперь это преобразование применяется к самому уравнению, а не к степенным суммам S_m . Фактически степенные суммы будут использованы только для определения r_0 .

В § 16 мы имели дело с уравнением шестой степени (16.4) и получили значение $r_0 = 1,29$, которое можно округлить до 1,3. Разделив

последовательные коэффициенты на последовательные степени 1, 3, мы получаем новое уравнение

$$u^6 - 0,38461u^5 + 0,22189u^4 - 0,41429u^3 + \\ + 0,86575u^2 - 0,05260u + 0,12645 = 0. \quad (1-18.1)$$

Мы знаем заранее, что по меньшей мере одна пара комплексных корней уравнения (1) будет лежать вблизи единичной окружности $|u| = 1$. Теперь мы используем известное преобразование, широко применяющееся в теории аналитических функций.

Простое обратное преобразование

$$\bar{z} = \frac{1}{z}, \quad (1-18.2)$$

рассматриваемое в комплексной плоскости, обладает замечательными свойствами. Оно дает «конформное отображение» плоскости на себя, преобразуя внешнюю область единичного круга во внутреннюю и обратно.

Более общее дробно-линейное преобразование в комплексной плоскости обладает следующим замечательным свойством: семейство окружностей и прямых (рассматриваемых, как окружности бесконечного радиуса) переводится им в семейство окружностей и прямых (рис. 1).

Рассмотрим дробно-линейное преобразование

$$u = \frac{1-v}{1+v}, \quad v = \frac{1-u}{1+u}. \quad (1-18.3)$$

Если комплексная переменная u перемещается вдоль единичной окружности, то соответствующая переменная v движется вдоль мнимой оси. Следовательно, корень на единичной окружности преобразуется в чисто мнимый корень.

Преобразование (3) легко может быть выполнено, если $f(u)$ есть полином относительно u степени n . Мы получаем тогда (после сокращения на множитель $(1+v)^{-n}$, не представляющий интереса для нашей цели) новый полином относительно v . Коэффициенты этого полинома зависят линейно от первоначальных коэффициентов a_k и могут быть получены путем умножения строки $a_0, a_1, a_2, \dots, a_n$ на определенную матрицу B_n^*) из $n+1$ строк и $n+1$ столбцов. Элементы этой матрицы для любых n могут быть заранее вычислены (см. табл. II приложения); они представляют собой целые числа. Например, для $n=1, 2$ и 3 мы получаем соответственно

$$\begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & -1 & 1 \\ -2 & 0 & 2 \\ 1 & 1 & 1 \end{bmatrix}, \quad \begin{bmatrix} -1 & 1 & -1 & 1 \\ 3 & -1 & -1 & 3 \\ -3 & -1 & 1 & 3 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

*) Об умножении матриц см. гл. II, § 3.

Эти матрицы легко могут быть составлены с помощью рекуррентного соотношения

$$b_{ik}^{(n+1)} = b_{(i-1)k}^{(n)} - b_{ik}^{(n)}. \quad (1-18.4)$$

Например, элемент второй строки и первого столбца (21) третьей матрицы получается вычитанием из элемента (11) (первая строка,

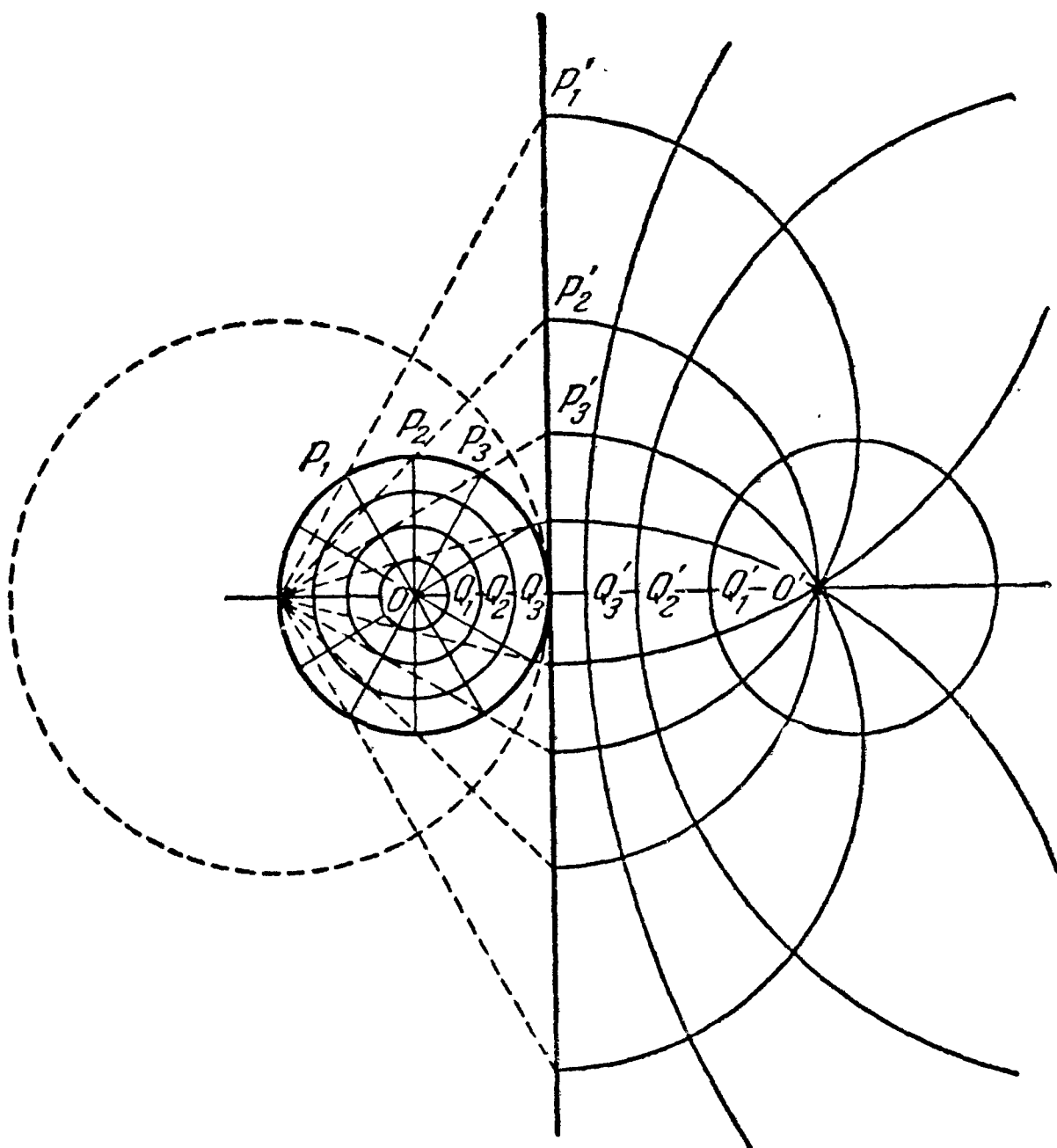


Рис. 1.

первый столбец) второй матрицы элемента (21) (вторая строка, первый столбец) той же матрицы: $1 - (-2) = 3$. Последний столбец, который следует добавить, всегда состоит из абсолютных значений соответствующих элементов первого столбца.

Интересным свойством матриц B является то, что их квадрат всегда пропорционален единичной матрице

$$B_n^2 = 2^n \cdot I \quad (1-18.5)$$

Это значит, что, «перемножая» какую-либо строку матрицы B_n на любой столбец той же матрицы, получаем 2^n по диагонали и 0 на всех других местах матрицы произведения. Например, в третьей матрице произведение второй строки на второй столбец дает

$$3 + 1 + 1 + 3 = 8,$$

а произведение второй строки на третий столбец —

$$-3 + 1 - 1 + 3 = 0.$$

Умножим теперь строку коэффициентов нашего уравнения (1)
1; — 0,38461; 0,22189; — 0,41429; 0,96575; — 0,05260; 0,12645
на матрицу B_6 табл. II. Умножение производится построчно. Мы получаем таким образом преобразованный ряд коэффициентов полинома шестой степени относительно v :

$$3,06379v^6 - 5,28162v^5 + 16,75769v^4 - 20,04644v^3 + \\ + 14,86053v^2 - 2,62554v + 1,36439 = 0 \quad (1-18.6)$$

(Численный контроль: сумма новых коэффициентов равна прежнему свободному члену, умноженному на 2^n *). Этот контроль абсолютно точен, так как умножение на целые числа не вводит ошибок округления. В нашем примере $\sum a'_i = 8,09280 = 64 \cdot 0,12645$.)

19. Корни, близкие к мнимой оси

Корень исходного уравнения относительно переменной u , который лежал близ единичной окружности, теперь переходит в корень относительно новой переменной v , который лежит около мнимой оси. Но если корень алгебраического уравнения с вещественными коэффициентами расположен около мнимой оси, то он легко может быть локализован. Допустим сначала, что этот корень лежит в точности на мнимой оси, т. е. равен iu ; тогда данное уравнение равносильно двум уравнениям, так как, очевидно, сумма членов с четными степенями (вещественная часть) и сумма членов с нечетными степенями (мнимая часть) в отдельности должны равняться нулю. Полагая

$$v^2 = -\lambda, \quad (1-19.1)$$

получаем для λ два алгебраических уравнения, степень которых не превышает *половины степени* исходного уравнения. Кроме того, для новых уравнений мы должны разыскивать лишь *вещественные* корни. Наша задача, таким образом, значительно упрощена.

*) Это свойство суммы коэффициентов $\sum_{k=0}^n a'_k$ легко вывести, если принять во внимание свойство элементов матрицы B_n , а именно: сумма элементов первых $n-1$ столбцов ее равна нулю, а сумма элементов последнего столбца равна 2^n .

Поэтому

$$a'_\alpha = \sum_i a_i b_{\alpha i}; \quad \sum_\alpha a'_\alpha = \sum_\alpha \left(\sum_i a_i b_{\alpha i} \right) = \sum_i \left(\sum_\alpha a_i b_{\alpha i} \right) = \sum_\alpha a_n b_{\alpha n} = a_n 2^n.$$

Мнимая часть исходного уравнения приводит к линейному уравнению, если оно было четвертой степени, и даже в случае, когда исходное уравнение имеет шестую степень, уравнение относительно λ все же будет только квадратным и может быть решено по обычным формулам. Нас при этом интересуют только положительные корни.

Мнимая часть (нечетные степени) численного примера (1-18.6) дает для λ :

$$5,28162\lambda^2 - 20,04644\lambda + 2,62554 = 0.$$

Это уравнение имеет два положительных корня:

$$\lambda_1 = 0,13583, \quad \lambda_2 = 3,65967.$$

Какой из них будет нас интересовать, заранее сказать нельзя. Мы должны подставить эти λ в вещественную часть исходного уравнения и посмотреть, является ли остаток малым или нет. В нашем случае оба λ_i дают малые остатки; это указывает на то, что обе пары комплексных корней лежат близ мнимой оси.

Дополнительное преимущество этого метода состоит в том, что поправка предварительного значения корня по схеме Ньютона (ср. § 5) оказывается очень упрощенной. «Подвижная полоса», используемая для вычисления $f(x_0)$, $f'(x_0)$ и $\frac{1}{2}f''(x_0)$, теперь состоит из чисел

$$-\lambda, 0, 1$$

в вертикальном расположении *). Ввиду наличия посредине нуля число умножений сокращается вдвое и вычисления существенно облегчаются.

Мы приведем схему подстановки для случая меньшего корня λ_1 :

$-0,13583$ 0 1	Частное	$f(v)$		Частное	$f'(v)$	
	3,06379	3,06379				
	—5,28162	—5,28162		18,38274	18,38274	
	16,34153	16,75769		—26,40810	—26,40810	
	—19,32904	—20,04644		—64,40810	67,03075	
↓	12,64086	14,86053	Остаток	—56,55231	—60,13932	Остаток
	0	—2,62554	—0,00008	0	29,72106	20,955543
	0	1,36439	—0,35262	0	—2,62554	5,05596
	$\frac{1}{2}f''(v)$					
	45,95685	45,95685		$v_0 = \sqrt{0,13583i} = 0,36855i,$		
	—52,81620	—52,81620		$f(v_0) = -0,3526,$		
	94,30380	100,54610		$f'(v_0) = 7,7231i + 5,0560,$		
	0	—60,13932	—52,96530	$\frac{1}{2}f''(v_0) = -19,5204i + 2,0512,$		
	0	14,86053	2,05124			

*). Предполагается, что подстановка производится по способу, указанному в § 10.

$$\begin{aligned}\frac{1}{h} &= \frac{5,0560 + 7,7231i}{0,3526} + \frac{2,0512 - 19,5204i}{5,0560 + 7,7231i} = \\ &= (14,338 + 21,902i) - (1,719 - 1,344i) = 12,619 + 20,558i, \\ h &= 0,02169 - 0,03533i, \\ v &= v_0 + h = 0,02169 + 0,33322i.\end{aligned}$$

Переходя обратно к первоначальной переменной u , получаем

$$\begin{aligned}u &= \frac{1-v}{1+v} = -1 + \frac{2}{1+v} = -1 + \frac{2}{1,02169 + 0,33322i} = \\ &= 0,76933 - 0,57706i.\end{aligned}$$

Наконец мы возвращаемся к исходному переменному x , умножая на 1,3:

$$\begin{aligned}x &= (0,76933 - 0,57706i) 1,3 = 1,0001 - 0,7502i \\ &\quad [\text{точно: } 1 - 0,75i].\end{aligned}$$

Схема введения поправки для бóльшего корня совершенно такая же. Однако целесообразно подставлять значение, *обратное* значению λ :

$$\bar{\lambda} = 0,27325, \quad \bar{v}_0 = \sqrt{0,27325i} = 0,52213i$$

в обращенный полином

$$1,36439 - 2,62554\bar{v} + 14,86053\bar{v}^2 - \dots + 3,06379\bar{v}^6.$$

Таким образом, мы получаем, если выполнить вычисления,

$$\begin{aligned}\bar{h} &= 0,02448 - 0,02291i, \\ \bar{v} &= 0,02448 + 0,49982i, \\ u &= \frac{\bar{v}-1}{\bar{v}+1} = 1 - \frac{2}{\bar{v}+1} = 1 - \frac{2}{0,02448 + 0,49982i} = \\ &= -0,57688 + 0,76932i, \\ x &= 1,3 (-0,57688 + 0,76932i) = -0,7499 + 1,0001i \\ &\quad [\text{точно: } -0,75 + i].\end{aligned}$$

20. Кратные корни

Всякий раз, когда значение $f(x_0)$ малó, метод Ньютона дает хорошую поправку при условии, однако, что $f'(x_0)$ в то же время не очень малó. Однако нужны особые предосторожности в том случае, когда окажется, что не только $f(x_0)$, но и $f'(x_0)$ малó, так что дробь $\frac{f(x_0)}{f'(x_0)}$ не обязательно будет малой. Малая величина $f'(x_0)$ указывает на то, что мы находимся вблизи двойного корня или вблизи двух очень близких корней. Отделение этих двух корней не может быть дости-

гнуто путем простой линейной интерполяции. Приходится вычислить $f(x_0)$, $f'(x_0)$ и $f''(x_0)$ и решить квадратное уравнение

$$\frac{1}{2} f''(x_0) h^2 + f'(x_0) h + f(x_0) = 0 \quad (1-20.1)$$

по обычной формуле. Полученные два корня, будучи добавлены к x_0 , дадут хорошие приближения для двух корней нашего уравнения. Приближенные значения корней x_1 и x_2 можно уточнить при помощи способа, описанного в § 5.

Аналогичная процедура потребуется в случае, когда не только $f(x_0)$ и $f'(x_0)$, но также $f''(x_0)$ так мало, что h , вычисленное из уравнения (1), не мало. В этом случае три корня очень близки друг другу. Для того чтобы их разделить, нам нужно вычислить еще $f'''(x_0)$ и решить кубическое уравнение

$$\frac{1}{6} f'''(x_0) h^3 + \frac{1}{2} f''(x_0) h^2 + f'(x_0) h + f(x_0) = 0. \quad (1-20.2)$$

Последующее применение схемы Ньютона отдельно к каждому полученному таким образом корню исправит, наконец, каждый корень до желаемой точности.

Разделение тесно сбившихся вместе кратных корней, как видим, представляет вообще громоздкий труд, который не может быть выполнен без большой вычислительной работы.

21. Алгебраические уравнения с комплексными коэффициентами

Пусть $F_n(x)$ есть полином с комплексными коэффициентами. Корни такого полинома комплексны, однако они не появляются здесь парами вида $\alpha \pm \beta i$. Метод степенных сумм остается пригодным, но вычисление моментов более высоких степеней очень замедляется комплексным характером коэффициентов. Произведение двух комплексных чисел требует четырех умножений, и поэтому время, нужное для получения радиуса максимальной окружности, намного возрастает.

Большой частью бывает выгоднее совсем избежать операций с комплексными числами при помощи умножения полинома $F_n(x)$ на сопряженный полином $F_n^*(x)$, где звездочка указывает на то, что в $F_n(x)$ каждое i заменяется на $-i$. Мы получаем таким образом полином степени $2n$:

$$f_{2n}(x) = F_n(x) F_n^*(x), \quad (1-21.1)$$

коэффициенты которого вещественны. Эта операция вводит n посторонних корней, так как к каждому корню $\alpha + i\beta$ добавляется дополнительный корень $\alpha - i\beta$. Но у нас появляется то большое преимущество, что устраняются операции с комплексными коэффициентами. Мы оперируем только с вещественным полиномом $f_{2n}(x)$ и получаем n пар сопряженных комплексных корней $x_k = \alpha_k \pm i\beta_k$.

Теперь надо решить, какой знак следует взять для мнимой части корня: плюс или минус. Это мы устанавливаем подстановкой x_k в первоначальный полином $F_n(x)$. Подвижная полоса опять содержит только вещественные коэффициенты*). Поэтому мы можем рассмотреть отдельно вещественную и мнимую части полинома $F_n(x)$ и провести процедуру независимо для каждой части, избегая комплексных чисел. Объединенные остатки получаются в виде

$$(a + ib)x_k + (c + id), \quad (1-21.2)$$

где коэффициенты a, b, c, d не малы. Приравняв этот остаток нулю, получим корень, который будет очень близок либо к $\alpha_k + i\beta_k$, либо к $\alpha_k - i\beta_k$.

Мы, таким образом, установим надлежащий знак и одновременно получим дополнительный контроль точности корня.

22. Анализ устойчивости

Всякий раз, когда механическая или электрическая система находится в состоянии равновесия, небольшие возмущающие силы будут стремиться нарушить это равновесие. Так как потенциальная энергия в окрестности равновесного состояния всегда есть положительно определенная форма относительно смещений, то система будет совершать небольшие колебания вокруг положения равновесия. Эти колебания можно рассматривать как линейную суперпозицию n «нормальных колебаний». Частоты этих колебаний определяются решением «характеристического уравнения», соответствующего динамической системе. В то время как силы трения вызывают затухание этих колебаний и возвращение системы к положению равновесия, положение изменяется, если в системе существует источник энергии, который может противодействовать этой тенденции сил трения и даже вызывать амплитуды все возрастающей величины. Общая проблема устойчивости приводит, таким образом, к решению алгебраического уравнения, комплексные корни которого могут выявить положительное или отрицательное торможение в зависимости от знака вещественных частей корней. Устойчивость системы требует, чтобы все корни характеристического уравнения лежали в отрицательной полуплоскости комплексной плоскости. Одного корня с положительной вещественной частью достаточно, чтобы была нарушена устойчивость системы.

При этих обстоятельствах представляет интерес решить, имеет ли заданное алгебраическое уравнение корни в положительной полуплоскости или нет, не входя в подробности точного определения корней. Английский физик Раус (E. J. Routh) нашел изящный метод испытания алгебраического уравнения на устойчивость, без какого-либо вы-

*) См. сноску на стр. 57.

числения его корней¹⁾. Однако иногда надо идти дальше и действительно локализовать корень неустойчивости, если система окажется неустойчивой. Для этой цели опять очень удобно использовать преобразование обратными радиусами. Заменим x на u и снова перейдем к переменной v с помощью преобразования (18.3). Это преобразование отображает всю левую полуплоскость $R(x) < 0$ в единичный круг. Следовательно, условие того, что все корни x_i лежат в отрицательной полуплоскости, равносильно условию, что все корни v_i по абсолютной величине меньше единицы. Поэтому радиус r_m максимальной окружности должен получиться меньшим 1, для того чтобы гарантировать устойчивость системы. Так как метод алгебраического деления дает возможность легко получить степенные суммы, а также r_m , то мы получаем при затрате сравнительно небольшого труда критерий устойчивости вместе с локализацией неустойчивого корня, если r_m оказывается бóльшим 1.

При преобразовании заданного уравнения важно, чтобы оно было дано в надлежащем масштабе. Мы нормируем масштаб требованием, чтобы свободный член уравнения был близок к 1 (в предположении, что коэффициент при x^n равен 1).

Пример 1. Корни рассмотренного ранее уравнения четвертой степени (14.5) имеют отрицательные вещественные части, и полином имеет, таким образом, «устойчивый» характер. Постараемся теперь установить этот факт, не вычисляя корней. Первоначальное уравнение есть

$$x^4 + 7,60x^3 + 23,34x^2 + 38,44x + 37,40 = 0. \quad (1-22.1)$$

Нормируем свободный член и переходим от x к u с помощью преобразования

$$x = -\frac{1}{0,4}u, \quad (1-22.2)$$

что сводится к умножению последовательных коэффициентов уравнения (1) на степени 0,4 и перемене знаков коэффициентов при нечетных степенях на обратные:

$$u^4 - 3,04u^3 + 3,7344u^2 - 2,4602u + 0,9574 = 0. \quad (1-22.3)$$

Новый свободный член близок к 1. Умножение на матрицу B_4 дает новое уравнение относительно v :

$$11,1920v^4 - 1,3300v^3 + 4,2756v^2 + 0,9892v + 0,1916 = 0. \quad (1-22.4)$$

Устойчивость этого уравнения может быть установлена с первого взгляда без дальнейших расчетов. Приняв, что v есть произвольное комплексное число, мы можем рассматривать члены алгебраического уравнения как векторы комплексной плоскости. Тот факт, что сумма

¹⁾ Описание метода Рауса см. в {2}, т. I, стр. 129; а также ссылку 3 в гл. II, стр. 154.

векторов равна нулю, геометрически означает, что векторный многоугольник замкнут, т. е. что конечная точка его совпадает с начальной точкой. Если мы теперь отделим первый вектор от остальных, то получим, что отрезок прямой от A до B дополняется до многоугольника ломаной линией, которая начинается в точке B и приходит в точку A . По свойству отрезков прямой линии вся длина пути от B до A никогда не может быть меньше, чем путь от A до B . Поэтому если мы обнаружим, что при некотором v имеет место неравенство

$$|a_1 v^{n-1}| + |a_2 v^{n-2}| + \dots < |a_0 v^n|, \quad (1-22.5)$$

то мы сможем быть уверены, что равенство (22.4) не удовлетворяется. Если мы разделим уравнение (22.4) на v^n и обозначим $\frac{1}{v}$ через $\bar{v}$, то сразу увидим, что

$$11,192 > 1,3300 |\bar{v}| + 4,2756 |\bar{v}|^2 + 0,9892 |\bar{v}|^3 + 0,1916 |\bar{v}|^4 \quad (1-22.6)$$

для любого $|\bar{v}| \leq 1$, т. е. для любого $|v| \geq 1$.

Вообще мы можем сказать, что если коэффициент при v^n больше, чем сумма абсолютных величин остальных коэффициентов, то уравнение относительно v не может иметь корня вне или на единичной окружности, что означает, что система устойчива. Этот простой критерий (который достаточен, но не необходим для устойчивости) в нашем случае удовлетворяется, и следовательно, устойчивость уравнения установлена.

Пример 2. Уравнение

$$x^6 + 4,24x^5 + 8,50x^4 + 10,27x^3 + 8,09x^2 + 3,96x + 0,98 = 0$$

также относится к стабильному типу. Свободный член здесь уже почти надлежащим образом нормирован, и мы можем после перемены законов у нечетных степеней сразу преобразовать уравнение к переменной v^6 . Умножение на матрицу B_6 дает

$$37,04v^6 - 2,06v^5 + 23,30v^4 + 1,24v^3 + 2,92v^2 + 0,18v + 0,10 = 0.$$

Здесь, как и в предыдущем примере, коэффициент при v^6 превышает сумму модулей остальных коэффициентов, что сразу указывает на устойчивость системы.

Пример 3. Предыдущий полином становится неустойчивым, если умножить его на множитель

$$x^2 - 0,6x + 1.$$

Это дает полином восьмой степени

$$x^8 + 3,64x^7 + 6,956x^6 + 9,41x^5 + 10,428x^4 + \\ + 9,376x^3 + 6,694x^2 + 3,372x + 0,98.$$

Преобразование к переменной v дает

$$51,856v^8 - 2,884v^7 + 128,924v^6 - 3,620v^5 + 64,668v^4 + \\ + 3,476v^3 + 7,732v^2 + 0,468v + 0,260 = 0.$$

Большие коэффициенты при v^4 , v^6 , v^8 позволяют тотчас же локализовать корни неустойчивости. Пренебрегая остальными членами, мы получаем следующее квадратное уравнение относительно $v^2 = -y$:

$$51,86y^2 - 128,92y + 64,67 = 0,$$

которое дает

$$y = 1,243 \pm 0,546.$$

Меньший корень меньше 1 и, таким образом, устойчив. Неустойчивый корень дает

$$v = \sqrt{-1,79} = \pm 1,33i.$$

Переходя обратно к первоначальной переменной x , найдем

$$-x = \frac{1-v}{1+v} = -1 + \frac{2}{1+v} = -0,28 \pm 0,96i, \\ x = 0,28 \mp 0,96i.$$

Эта весьма близкая оценка истинного значения корня неустойчивости, который равен $0,3 \pm 0,95i$.

Пример 4. Не всегда условия для выделения корней неустойчивости столь благоприятны, как в примере 3. В примере, приведенном у Догерти — Келлера, [{2}, I, стр. 129], нужно исследовать на устойчивость следующее уравнение шестой степени:

$$x^6 + 68,6x^5 + 785x^4 + 7213x^3 + 50\,700x^2 + 8200x + 43500 = 0.$$

Для приближенного нормирования свободного члена мы полагаем

$$x = -10u$$

и получаем

$$u^6 - 6,86u^5 + 7,85u^4 - 7,213u^3 + 5,07u^2 - 0,082u + 0,435 = 0.$$

Преобразование к переменной v дает

$$28,51v^6 - 36,062v^5 + \\ + 21,676v^4 - 0,18v^3 - 4,466v^2 + 18,162v + 0,2 = 0.$$

Прежний критерий устойчивости здесь неприменим, так как сумма абсолютных величин последних шести коэффициентов превышает первый коэффициент. Следовательно устойчивый или неустойчивый характер уравнения еще невыяснен, и нам надо вычислить несколько степенных сумм. Разделив на коэффициент при v^6 и ограничиваясь двумя десятичными знаками, получаем

$$v^6 - 1,26v^5 + 0,76v^4 - 0,01v^3 - 0,16v^2 + 0,64v = 0.$$

Уравнение может быть, следовательно, приведено к уравнению пятой степени, так как корень $v=0$ (строго говоря, v — очень малое) не представляет интереса. Алгебраическое деление по схеме подвижной полосы дает последовательные степенные суммы:

$k =$	0	$S_k =$	5
	1		1,26
	2		0,0676
	3		— 0,87242
	$\vdots$		$\vdots$
	10		3,7941
	11		5,5555
	12		6,0846

Как только мы дошли до степенной суммы, превышающей 5, можно остановиться, ибо становится ясно, что система неустойчива. Действительно, последовательное возвышение в степень постоянно уменьшает абсолютную величину корней, если все корни лежат внутри единичной окружности. Следовательно, невозможно, чтобы сумма $v_1^k + v_2^k + \dots + v_n^k$ когда-либо превысила n , если все $|v_i|$ меньше 1. Если нашей целью является просто установить неустойчивость системы, то цель уже достигнута при получении $S_{11} > 5$. Но если нужно локализовать корень неустойчивости, то мы можем вычислить еще несколько сумм S_m и получить этот корень, решив квадратное уравнение (15.14).

Этот метод установления устойчивости системы, очевидно, перестает быть действительным, если критическое значение r очень близко к 1. Но это означает, что корень в первоначальной переменной u очень близок к мнимой оси. Так как корни, близкие к мнимой оси, как мы видели, легко могут быть локализованы (путем разыскания вещественных корней вспомогательной системы двух уравнений), то весьма слабо неустойчивая система так или иначе может быть выявлена и не нуждается в дальнейшем исследовании.

ЛИТЕРАТУРА К ГЛАВЕ I

- [1] См. {11}, гл. IX.
- [2] См. {14}, гл. VI.
- [3] Бохер М., Введение в высшую алгебру, ОНТИ, М.-Л., 1933.
- [3а] Загускин В. Л., Справочник по численным методам решения алгебраических и трансцендентных уравнений, Физматгиз, М., 1960.

Статья:

- [4] Fry Th. C., Some Numerical Methods of Locating Roots of Polynomials, Quart. Applied. Math. 3, 89 (1945).

ГЛАВА II

МАТРИЦЫ И ПРОБЛЕМЫ СОБСТВЕННЫХ ЗНАЧЕНИЙ

1. Исторический обзор

Решение систем линейных алгебраических уравнений имеет очень давнюю историю. Индийцы, которые являются изобретателями десятичной системы счисления и алгебраического метода, решали системы линейных алгебраических уравнений, начиная с VI века. На первых этапах развития алгебры символами обозначали только неизвестные, тогда как заданные коэффициенты вводили в систему в числовом виде. Полная символизация алгебры наступила во времена знаменитого французского алгебраиста Франсуа Виета (Viète, 1540—1603). Эта символизация дала возможность развить общие методы решения системы линейных уравнений. Великий философ и математик Лейбниц изобрел понятие определителя и ввел для него обозначение, которым мы по существу пользуемся по сей день. Он показал, что решение всякой совместной системы линейных алгебраических уравнений можно получить, составив отношение двух определителей.

В течение XIX века операторная точка зрения сделалась центром внимания. В то время как в арифметике важен лишь конечный численный ответ, для алгебры существенны *операции*, производимые при получении результатов. Нас интересует не конечный численный ответ задачи, а операции, с помощью которых этот ответ может быть получен. Поэтому в чисто алгебраической задаче раскрывают скобки, освобождаются от дробей, производят умножение и т. д. чисто *символически*. Мы не заботимся о том, каково численное значение символов, входящих в задачу. Наши заключения основываются на том факте, что все числа удовлетворяют некоторым общим законам, так называемым «постулатам алгебры». Это дает возможность делать правильные заключения, независимо от того, каковы фактические численные параметры задачи. Мы должны выполнять сложные алгебраические операции не над заданными числами, а только над *символами*, имитирующими свойства чисел. Числовые значения символов используются только в конечной формуле, которая указывает, какие операции нужно выполнить над заданными числами, чтобы получить ответ.

Основные арифметические действия — сложение и вычитание, умножение и деление — первоначально выполнялись только над натуральными числами; начав с целых чисел, постепенно расширяли числовую область, вводя отрицательные числа и дроби. В алгебре добавились два новых основных действия — возвышение в степень и извлечение корня. Эти действия вызвали дальнейшее расширение числовой области: были введены иррациональные числа и комплексные числа вида $a + bi$. Все эти числа удовлетворяют основным постулатам алгебры и, таким образом, алгебраические действия над ними выполняются одинаково. В 1859 английский математик Кели значительно расширил область алгебры, показав, что и «матрица», хотя она и состоит из *системы* величин, может быть рассматриваема как единый алгебраический символ, который удовлетворяет всем постулатам обычной алгебры, за исключением переместительного закона умножения. Таким образом, алгебра матриц представляет собой пример некоммутативной алгебры. Матричная алгебра отличается, однако, еще одной чертой от обычной алгебры. Каждая матрица удовлетворяет своему собственному характеристическому уравнению и приводит поэтому к полиномиальному тождеству, не имеющему аналога в алгебре обыкновенных вещественных или комплексных чисел *). Эта сторона теории матриц была развита Сильвестром (1851) (он же ввел термин «собственные корни» (latent roots)) и позднее Вейерштрассом (1868). Наиболее полная алгебраическая теория характеристического уравнения была окончательно дана Фробениусом (1879).

Фредгольм (Fredholm, 1900) ввел понятие матриц бесконечного порядка и расширил алгебраическую теорию характеристического уравнения на случай *бесконечно многих* переменных. Эта теория стала основанием теории систем ортогональных функций (ср. V, 16) и геометрической трактовки линейных дифференциальных и интегральных операторов.

Область приложений. Характеристическое уравнение вместе с его собственными значениями и соответствующими собственными векторами играет основную роль в теории колебаний, будь то колебания механического или электрического, макроскопического или микроскопического характера. Упругие колебания моста или другого жесткого сооружения, неустойчивые колебания крыла самолета, неустановившиеся колебания электрической сети, колебания атомов и молекул в волновой механике — все это примеры применения характеристического уравнения. Явление выпучивания упругой системы также является проблемой собственных значений, так как выпучивание возникает, когда наименьшая частота колебаний упругой системы достигает нулевого значения.

В волновой механике Шредингера атомарные и молекулярные колебания играют фундаментальную роль в описании физических и

*) Подробно об этом см. в § 5 этой главы.

химических свойств материи. Собственные значения волнового уравнения Шредингера прямо пропорциональны значениям энергии различных квантовых состояний.

Разложение решения дифференциального уравнения в частных производных, соответствующего заданным граничным условиям, по ортогональным функциям, ассоциированным с данным дифференциальным оператором, дает мощное средство для представления решения через заданные граничные значения. Функция Грина, соответствующая данному дифференциальному оператору, часто недоступна в явной форме. Билинейное разложение дает косвенный метод образования функции Грина.

Теория линейных дифференциальных и интегральных операторов очень выигрывает в ясности и сжатости при введении соответствующих собственных функций как вспомогательной базисной системы. В геометрической интерпретации это означает, что соответствующая поверхность второго порядка преобразуется к своим главным осям. Преобразование квадратичных форм к главным осям становится, таким образом, фундаментальным связующим звеном между весьма различными ветвями математики. Решение системы линейных алгебраических уравнений, матричное исчисление, общая теория линейных дифференциальных и интегральных операторов — все эти проблемы могут рассматриваться как различные формулировки одной и той же основной проблемы, а именно преобразование к главным осям поверхности второго порядка в евклидовом пространстве конечного или бесконечного числа измерений.

2. Векторы и тензоры

Векторы отличаются от скаляров тем, что имеют не только величину, но и *направление*. Но они являются не единственными величинами, для изучения которых приходится выйти за область скаляров. Векторы можно изучать в какой-либо системе координат и задавать с помощью их координат. Так, например, в пространстве трех измерений вектор a имеет координаты a_1, a_2, a_3 и изображается формулой

$$a = a_1 i + a_2 j + a_3 k,$$

или, в иной форме,

$$a = (a_1, a_2, a_3).$$

Эта запись является кратким указанием на то, что три числа a_1, a_2, a_3 , называемые координатами вектора, однозначно характеризуют вектор a *).

*) Если, конечно, заданы еще и координатные векторы i, j, k . Поэтому указанное выше обозначение $a = (a_1, a_2, a_3)$ является неполным и его следовало бы заменить, например, таким: $a = (a_1, a_2, a_3)i, j, k$. Этого, однако, не делают в целях краткости.

В процессе развития естественных наук в XIX веке было обнаружено, что векторы представляют собой лишь очень узкий класс не скалярных величин, так называемых «тензоров». Векторы являются простейшим классом тензоров, а именно тех тензоров, которые могут быть охарактеризованы системой чисел a_i с одним индексом. В общем случае индексов может быть два, три или более: a_{ij} , a_{ijk} ...

Среди этих тензоров наиболее важны тензоры *второго ранга**). Тензор второго ранга характеризуется двумя индексами: a_{ik} . Составляющие (или координаты) такого тензора могут быть расположены в следующей двумерной квадратной таблице:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{pmatrix}.$$

Таким образом, первый индекс указывает *строку*, а второй индекс — *столбец*, которому принадлежит составляющая a_{ik} . Заклячая схему в скобки, мы хотим показать, что *всю совокупность* этих составляющих мы должны рассматривать как одно целое. Ни одна из этих составляющих не имеет самостоятельного значения. Только совокупность всех составляющих определяет тензор подобно тому, как только вся совокупность координат образует вектор**).

3. Матрицы как алгебраические объекты

Кели (Cauley) сделал фундаментальное открытие, что так расположенную систему чисел можно рассматривать как единый алгебраический объект A , над которым могут производиться определенные алгебраические действия. Мы можем складывать, вычитать, умножать и делить так расположенные системы чисел аналогично тому, как мы поступаем с алгебраическими величинами. Такую систему чисел мы будем называть «матрицей» и обозначать ее через

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}; \text{ а также } A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}.$$

*) В математической литературе принято также говорить «второй валентности» или второго порядка.

**) Существенно добавить, что эта совокупность составляющих определяет тензор относительно некоторой выбранной системы координат, заданной векторами, например, i, j, k для трехмерного пространства. Поэтому указанное выше обозначение тензора с помощью матрицы (a_{ik}) является неполным и его следовало бы заменить, например, таким: $(a_{ik})_{i, j, k}$. При изменении ко-

Числа a_{ik} называются элементами матрицы A . В отличие от определителя

$$|A| = \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix}, \quad (2-3.1)$$

мы не считаем матрицу A одним числом, которое получается с помощью определенного процесса над элементами a_{ik} матрицы. Матрица A означает просто *всю совокупность в целом* чисел a_{ik} , но расположенных определенным образом в строки и столбцы. Каждый элемент a_{ik} имеет свое определенное место в расположении и не может быть замещен другим элементом. Следовательно, матрица определяется как совокупность чисел, расположенных по строго установленной геометрической схеме.

Когда мы говорим, что можем производить алгебраические действия над матрицами, то это означает, что основные действия алгебры, а именно сложение, вычитание, умножение и деление, возвышение в степень и извлечение корня, могут быть целесообразно обобщены на область матриц.

Два основных действия, из которых остальные могут быть выведены, суть сложение и умножение. Вычитание есть действие, обратное сложению, а деление есть действие, обратное умножению. Поэтому достаточно знать, как складывать и как умножать матрицы.

Сложение матриц выполняется очень просто. Если заданы две матрицы

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}, \quad B = \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \dots & \dots & \dots & \dots \\ b_{n1} & b_{n2} & \dots & b_{nn} \end{bmatrix}, \quad (2-3.2)$$

то мы складываем соответствующие элементы и получаем новую матрицу

$$C = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \dots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \dots & a_{2n} + b_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} + b_{n1} & a_{n2} + b_{n2} & \dots & a_{nn} + b_{nn} \end{bmatrix}. \quad (2-3.3)$$

Эту новую матрицу C мы называем суммой матриц A и B :

$$C = A + B. \quad (2-3.4)$$

ординатной системы, т. е. при переходе к другим координатным векторам, скажем i', j', k' , этот же тензор определялся бы новой совокупностью чисел a'_{ik} и обозначался бы так: $(a'_{ik})_{i', j', k'}$.

Рассматривая x как заданный вектор мы можем сказать, что матрица A ставит в соответствие данному вектору x новый вектор b . Мы можем также сказать, что матрица A «преобразует» вектор x в новый вектор b . Рассмотрим, например, поворот вектора x вокруг некоторой оси на какой-либо угол. Такое вращение представляет собой пример умножения вектора x на некоторую матрицу. Однако матрица, соответствующая вращению твердого тела в пространстве, является весьма *специальной* матрицей. Матрица общего вида преобразует вектор x в новый вектор b , но это преобразование не представляет собой простого вращения.

Допустим теперь, что мы исходим из вектора u и преобразуем его в вектор v с помощью матрицы A . Затем мы повторяем этот процесс и преобразуем v в новый вектор w с помощью другой матрицы B :

$$v = Au, \quad w = Bv. \quad (2-3.8)$$

Итак, вектор w получен из v , но вектор v сам получен из u . Следовательно, опуская промежуточный вектор v , мы можем сказать, что w получен из u *). Если мы в предыдущее равенство подставим вместо v его выражение через u , то получим

$$w = BAu. \quad (2-3.9)$$

Поэтому прямой переход от u к w можно представить в виде

$$w = Cu, \quad (2-3.10)$$

причем для матрицы C оказывается целесообразней следующая запись:

$$C = BA. \quad (2-3.11)$$

Это приводит к правилу, по которому получается произведение двух матриц B и A . Если мы выполним указанную подстановку, то найдем, что произвольный элемент c_{ik} матрицы C произведения должен строиться следующим образом. Выбираем i -ю строку первого множителя B и k -й столбец второго множителя A (рис. 3). «Перемножаем» эти два ряда. «Умножение» здесь обозначает перемножение соответствующих элементов этих рядов и последующее суммирование полученных произведений. Например, если

$$\begin{aligned} v_1 &= 3u_1 - 2u_2, & w_1 &= v_1 + v_2, \\ v_2 &= -u_1 + 4u_2, & w_2 &= 2v_1 - 5v_2, \end{aligned}$$

то

$$\begin{aligned} w_1 &= 3u_1 - 2u_2 - u_1 + 4u_2 = 2u_1 + 2u_2, \\ w_2 &= 6u_1 - 4u_2 + 5u_1 - 20u_2 = 11u_1 - 24u_2. \end{aligned}$$

*) Легко понять, что переход от u к w осуществляется также с помощью линейных формул.

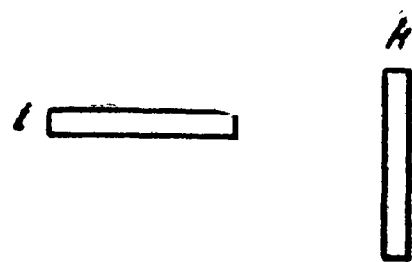


Рис. 3.

Таким образом,

$$\begin{bmatrix} 1 & 1 \\ 2 & -5 \end{bmatrix} \cdot \begin{bmatrix} 3 & -2 \\ -1 & 4 \end{bmatrix} = \\ = \begin{bmatrix} 1 \cdot 3 + 1(-1) = 2 & 1(-2) + 1 \cdot 4 = 2 \\ 2 \cdot 3 + (-5)(-1) = 11 & 2(-2) + (-5)4 = -24 \end{bmatrix}.$$

Если такое же «перемножение» произвести в *обратной* последовательности, т. е. изменив порядок следования перемножаемых матриц, то мы *не* получим прежнего результата. В самом деле, если

$$\begin{aligned} v_1 &= u_1 + u_2, & w_1 &= 3v_1 - 2v_2, \\ v_2 &= 2u_1 - 5u_2, & w_2 &= -v_1 + 4v_2, \end{aligned}$$

то

$$\begin{aligned} w_1 &= 3u_1 + 3u_2 - 4u_1 + 10u_2 = -u_1 + 13u_2, \\ w_2 &= -u_1 - u_2 + 8u_1 - 20u_2 = 7u_1 - 21u_2, \\ \begin{bmatrix} 3 & -2 \\ -1 & 4 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 \\ 2 & -5 \end{bmatrix} &= \\ &= \begin{bmatrix} 3 \cdot 1 + (-2)2 = -1 & 3 \cdot 1 + (-2)(-5) = 13 \\ (-1)1 + 4 \cdot 2 = 7 & (-1)1 + 4(-5) = -21 \end{bmatrix}. \end{aligned}$$

Это показывает, что произведения AB и BA отличаются одно от другого. Обычный закон коммутативности умножения не выполняется для матриц.

Но в то же время закон *ассоциативности* умножения выполняется:

$$A(BC) = (AB)C. \quad (2-3.12)$$

Действительно, преобразуем u в v , затем в w и, наконец, в z с помощью следующих операций:

$$v = Cu, \quad w = Bv, \quad z = Aw.$$

Тогда

$$w = B(Cu) = BCu,$$

и поэтому

$$z = Aw = A\{(BC)u\} = \{A(BC)\}u. \quad (\alpha)$$

С другой стороны,

$$z = A(Bv) = (AB)v,$$

и поэтому

$$z = AB(Cu) = \{(AB)C\}u. \quad (\beta)$$

Сопоставляя (α) и (β) и учитывая, что u — произвольный вектор, получим

$$A(BC) = (AB)C.$$

нулевыми правыми частями) линейных уравнений относительно $x_1, \dots, x_n$:

$$\left. \begin{aligned} (a_{11} - \lambda)x_1 + a_{12}x_2 + \dots + a_{1n}x_n &= 0, \\ a_{21}x_1 + (a_{22} - \lambda)x_2 + \dots + a_{2n}x_n &= 0, \\ \dots &\dots \\ a_{n1}x_1 + a_{n2}x_2 + \dots + (a_{nn} - \lambda)x_n &= 0. \end{aligned} \right\} \quad (2-4.3)$$

Матрицей коэффициентов этой системы уравнений служит в основном все та же первоначальная матрица A , в которой, однако, из всех диагональных элементов вычитается λ .

Как известно, система n однородных линейных уравнений с n неизвестными имеет «ненулевое» решение (т. е. такое, в котором хотя бы одно из чисел x_i не равно нулю)*), лишь при выполнении вполне определенного условия, а именно, что *определитель* этой системы равен нулю; в нашем случае

$$\begin{vmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - \lambda & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} - \lambda \end{vmatrix} = 0. \quad (2-4.4)$$

Разлагая определитель, стоящий слева, по правилам, известным из высшей алгебры, мы, очевидно, получим полином n -й степени относительно числа λ .

Таким образом, уравнение (1) только в том случае может иметь своим решением отличный от нуля вектор x , если множитель пропорциональности λ равен одному из корней уравнения (4).

Непосредственное разложение определителя (4) по степеням λ представляет собой весьма кропотливую работу, если $n > 4$. Существуют другие, менее прямые, но вычислительно более простые методы определения числа λ , удовлетворяющие условию (4). Однако для теоретических целей само существование условия (4) является исключительно важным. Целесообразно умножить определитель (4) на $(-1)^n$. Тогда старший член полинома принимает вид λ^n :

$$\begin{aligned} (-1)^n \begin{vmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - \lambda & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} - \lambda \end{vmatrix} = \\ = \lambda^n + C_{n-1}\lambda^{n-1} + C_{n-2}\lambda^{n-2} + \dots + C_0. \end{aligned} \quad (2-4.5)$$

Пусть, например, $n=3$. Вычисляя определитель (5), получим

$$-\begin{vmatrix} 2 - \lambda & 0 & 3 \\ 1 & -1 - \lambda & 5 \\ 0 & 4 & -2 - \lambda \end{vmatrix} = \lambda^3 + \lambda^2 - 24\lambda + 24.$$

*) Всякая система однородных линейных уравнений имеет, конечно, «нулевое» решение, но нас интересует вектор, имеющий направление, т. е. вектор, у которого хотя бы одна из координат отлична от нуля.

комбинацией первых *двух* собственных векторов

$$x = \alpha_1 u_1 + \alpha_2 u_2^*).$$

Аналогично уравнение

$$(A - \lambda_1)(A - \lambda_2)(A - \lambda_3)x = 0$$

будет удовлетворяться произвольной линейной комбинацией первых *трех* собственных векторов

$$x = \alpha_1 u_1 + \alpha_2 u_2 + \alpha_3 u_3.$$

Наконец полное уравнение, которое содержит все корневые множители:

$$(A - \lambda_1)(A - \lambda_2)(A - \lambda_3) \dots (A - \lambda_n)x = 0,$$

будет удовлетворяться произвольной линейной комбинацией *всех* n собственных векторов

$$x = \alpha_1 u_1 + \alpha_2 u_2 + \dots + \alpha_n u_n.$$

Но n собственных векторов $u_1, u_2, \dots, u_n$ охватывают *все пространство*¹⁾, и таким образом, x оказывается *произвольным вектором* n -мерного пространства. Это означает, что матрица

$$H = (A - \lambda_1)(A - \lambda_2)(A - \lambda_3) \dots (A - \lambda_n), \quad (2-5.1)$$

действуя на *произвольный* вектор x , дает нуль:

$$Hx = 0. \quad (2-5.2)$$

Полученное тождество может выполняться только при условии, что матрица H *тождественно равна нулю*. Мы, следовательно, нашли, что произвольная матрица удовлетворяет следующему полиномиальному тождеству:

$$(A - \lambda_1)(A - \lambda_2) \dots (A - \lambda_n) \equiv 0. \quad (2-5.3)$$

Перепишем это тождество в более полном виде ^{**) :}

$$(A - \lambda_1 I)(A - \lambda_2 I) \dots (A - \lambda_n I) \equiv 0. \quad (2-5.4)$$

^{*)} В самом деле,

$(A - \lambda_1)(A - \lambda_2)(\alpha_1 u_1 + \alpha_2 u_2) = \alpha_1 (A - \lambda_1)(A - \lambda_2)u_1 + \alpha_2 (A - \lambda_1)(A - \lambda_2)u_2 =$
 $=$ (если учесть, что $(A - \lambda_2)u_2 = 0$) $= \alpha_1 (A - \lambda_1)(A - \lambda_2)u_1$. Матрицы $(A - \lambda_1)$ и $(A - \lambda_2)$ коммутативны, ибо, вычисляя по правилам матричной алгебры каждое из произведений $(A - \lambda_1)(A - \lambda_2)$ и $(A - \lambda_2)(A - \lambda_1)$, получим один и тот же результат: $AA - (\lambda_1 + \lambda_2)A + \lambda_1 \lambda_2$. Поэтому

$$\alpha_1 (A - \lambda_1)(A - \lambda_2)u_1 = \alpha_1 (A - \lambda_2)(A - \lambda_1)u_1 = 0.$$

¹⁾ Это неверно для «неполноосных» матриц (ср. § 11), но тогда теорема Гамильтона — Кели устанавливается путем предельного перехода.

^{**)} См. примечание на стр. 76.

Если мы образуем квадрат и куб матрицы A согласно правилам матричного умножения, то получим

$$A^2 = \begin{bmatrix} 4 & 12 & 0 \\ 1 & 21 & -12 \\ 4 & -12 & 24 \end{bmatrix}, \quad A^3 = \begin{bmatrix} 20 & -12 & 72 \\ 23 & -69 & 132 \\ -4 & 108 & -96 \end{bmatrix}.$$

Теперь образуем выражение $A^3 + A^2 - 24A + 24I$:

$$\begin{bmatrix} 20 + 4 - 48 + 24 = 0, & -12 + 12 - 0 + 0 = 0, \\ 23 + 1 - 24 + 0 = 0, & -69 + 21 + 24 + 24 = 0, \\ -4 + 4 - 0 + 0 = 0, & 108 - 12 - 96 + 0 = 0, \\ & 72 + 0 - 72 + 0 = 0 \\ & 132 - 12 - 120 + 0 = 0 \\ & -96 + 24 + 48 + 24 = 0 \end{bmatrix}.$$

Мы, следовательно, убедились в том, что

$$A^3 + A^2 - 24A + 24A^0 \equiv 0.$$

Существование полиномиального соотношения вида

$$A^n + c_{n-1}A^{n-1} + \dots + c_1A + c_0A^0 = 0 \quad (\alpha)$$

является причиной принципиального отличия матричной алгебры от обычной алгебры чисел даже в случае операций с одной-единственной матрицей. Если x есть обыкновенная алгебраическая величина, то мы можем составить ее первую, вторую, третью и т. д. степени. Все эти степени x будут линейно независимы друг от друга, т. е. никакая степень не может быть представлена как линейная комбинация более низких степеней (хотя в заданном конечном интервале такое приведение возможно с высокой степенью точности, ср. VII, 9). Для матриц с n строками и n столбцами положение другое. В силу (α) A^n может быть представлено в виде линейной комбинации более низких степеней матрицы A^*). Этот же факт останется верным и для A^{n+1} , A^{n+2} , ... и вообще для A^{n+k} ($k=0, 1, 2, \dots$). Следовательно, любой полином относительно A , имеющий степень, большую $n-1$, может быть всегда сведен к полиному, степень которого не превышает $n-1$.

*) Следует, впрочем, заметить, что зависимость между степенями матрицы A вида

$$A^n + c_{n-1}A^{n-1} + \dots + c_0I \equiv 0$$

не вполне аналогична линейной зависимости между степенями алгебраического числа x :

$$x^n + k_{n-1}x^{n-1} + \dots + k_0 \equiv 0.$$

В последнем соотношении предполагается, что коэффициенты $k_{n-1}, k_{n-2}, \dots, k_0$ не зависят от переменной x , тогда как в равенстве (α) коэффициенты $c_{n-1}, c_{n-2}, \dots, c_0$ определенным образом связаны с элементами матрицы A .

Другое важное следствие из существования тождества Гамильтона — Кели мы получим при рассмотрении вопроса об операции деления на матрицу.

В алгебре чисел операция деления единицы на заданное число x возможна для любого числа x , не равного нулю.

В матричной же алгебре операция деления единичной матрицы I на матрицу A может оказаться невозможной и в том случае, когда матрица A не равна нулю, т. е. и тогда, когда не все ее элементы равны нулю.

Легко убедиться (используя тождество Гамильтона — Кели), что деление единичной матрицы на матрицу A возможно в том и только в том случае, когда ни одно из собственных значений матрицы не равно нулю. В самом деле, переписав тождество (α) в виде

$$A(A^{n-1} + c_1 A^{n-2} + \dots + c_{n-2} A^1 + c_{n-1} A^0) + c_n I = 0 \quad (\beta)$$

и принимая во внимание соотношение

$$c_n = \lambda_1 \lambda_2 \dots \lambda_n,$$

придем к выводу, что в случае, когда ни одно из собственных значений λ_i не равно нулю, имеет место равенство

$$\left\{ -\frac{1}{c_n} (A^{n-1} + c_1 A^{n-2} + \dots + c_{n-2} A^1 + c_{n-1} A^0) \right\} A = I.$$

Таким образом, в рассматриваемом случае существует матрица, которая, будучи умножена на заданную матрицу A , дает единичную матрицу I .

Эту матрицу целесообразно считать результатом деления единичной матрицы I на матрицу A и обозначить ее $\frac{I}{A}$ или, короче, A^{-1} . Таким образом, в случае, когда ни одно из собственных значений матрицы A не равно нулю, матрица A^{-1} существует и определяется по формуле

$$A^{-1} = -\frac{1}{c_n} (A^{n-1} + c_1 A^{n-2} + \dots + c_{n-2} A^1 + c_{n-1} A^0).$$

Остается заметить, что в случае, когда хотя бы одно из собственных значений матрицы A равно нулю, коэффициент c_n становится также равным нулю. Тогда тождество (β) приобретает вид

$$A(A^{n-1} + c_1 A^{n-2} + \dots + c_{n-2} A^1 + c_{n-1} A^0) = 0,$$

и, следовательно, матрица A , не равная нулю, представляет собой делитель нуля. В этом случае не существует такой матрицы K , для которой $AK = I$.

Задача нахождения обратной матрицы (задача «обращения» матрицы) имеет фундаментальное значение при решении системы линейных уравнений. Мы видим, что эта задача не имеет решения, если

матрица A имеет нулевое собственное значение. В строго математическом смысле проблема обращения неразрешима лишь в том случае, когда какое-либо из собственных значений матрицы A *в точности* равно нулю. Но с практической точки зрения мы встречаемся с большими вычислительными трудностями не только когда A имеет нулевое собственное значение, но и тогда, когда A имеет одно или несколько *весьма малых собственных значений*. Математический анализ таких «почти особых систем» заслуживает особо пристального внимания. Строго говоря, проблема обращения матрицы может быть рассмотрена без каких-либо сведений относительно проблемы собственных значений. Но в действительности мы не можем понять своеобразных особенностей особой или почти особой системы, если рассматривать эту проблему в отрыве от проблемы собственных значений матрицы.

Проблема собственных значений играет весьма важную роль во всех явлениях неустойчивых колебаний и вибраций, так как частота упругих или электрических колебаний определяется собственными значениями некоторой матрицы, тогда как собственные векторы, или главные оси, этой матрицы указывают формы этих колебаний. Но даже исследования чисто статических явлений, таких, как анализ устойчивости самолета или проблема изгиба, эквивалентны проблеме собственных значений. Анализ собственных значений матриц становится, таким образом, ведущей темой технических исследований наших дней.

6. Численный пример полного анализа собственных значений

Представляет интерес проведение до конца на некотором численном примере всех операций, которые ведут к полному анализу собственных значений заданной матрицы. Мы выберем поэтому особенно простую матрицу только из трех строк и столбцов, для того чтобы свести до минимума численные подсчеты и все же выявить все характерные черты задачи нахождения собственных значений и собственных векторов.

Пусть дана следующая матрица:

$$A = \begin{bmatrix} 33 & 16 & 72 \\ -24 & -10 & -57 \\ -8 & 4 & -17 \end{bmatrix}. \quad (2-6.1)$$

Нашей целью будет получить все три собственных значения и три собственных вектора, принадлежащих этой матрице.

Прежде всего мы составим характеристический полином, вычтем из диагональных элементов λ , вычислим по правилам высшей алгебры определитель полученной матрицы $A - \lambda I$ (например, путем его разложения по элементам первой строки) и умножим полученный полином третьей степени относительно λ на $(-1)^3$ — для того чтобы в харак-

теристическом уравнении коэффициент перед λ^3 был положительным:

$$\begin{aligned}
 (-1)^3 & \begin{vmatrix} 33 - \lambda & 16 & 72 \\ -24 & -10 - \lambda & -57 \\ -8 & -4 & -17 - \lambda \end{vmatrix} = \\
 & = \begin{vmatrix} \lambda - 33 & -16 & -72 \\ -24 & -10 - \lambda & -57 \\ -8 & -4 & -17 - \lambda \end{vmatrix} = \\
 & = (\lambda - 33)[(10 + \lambda)(17 + \lambda) - 228] + 16[24(17 + \lambda) - 456] - \\
 & - 72[96 - 8(10 + \lambda)] = (\lambda - 33)(\lambda^2 + 27\lambda - 58) + 16(24\lambda - 48) - \\
 & - 72(-8\lambda + 16) = \lambda^3 + 27\lambda^2 - 58\lambda - \\
 & \quad - 33\lambda^2 - 891\lambda + 1914 + \\
 & \quad + 384\lambda - 768 + \\
 & \quad + 576\lambda - 1152 \\
 & \hline
 & \lambda^3 - 6\lambda^2 + 11\lambda - 6.
 \end{aligned} \tag{2-6.2}$$

Таким образом, характеристическое уравнение имеет вид

$$\lambda^3 - 6\lambda^2 + 11\lambda - 6 = 0. \tag{2-6.3}$$

Оно имеет корни

$$\lambda_1 = 1, \quad \lambda_2 = 2, \quad \lambda_3 = 3. \tag{2-6.4}$$

Мы проверим теперь тождество Гамильтона — Кели, — оно получается из характеристического уравнения, если заменить в нем λ через A :

$$A^3 - 6A^2 + 11A - 6I = 0, \tag{2-6.5}$$

или, в иной форме,

$$A^3 = 6A^2 - 11A + 6I. \tag{2-6.6}$$

Для того чтобы проверить это равенство, нужно составить квадрат и куб первоначальной матрицы. Для умножения двух матриц надо, как было уже сказано выше, «помножить» *строки* первой матрицы на столбцы второй матрицы. При этих условиях нелегко поставить в правильное соответствие элементы перемножаемых строки и столбца и легко можно ошибиться. Полезно поэтому предварительно «транспонировать» первую матрицу, т. е. сделать ее строки столбцами; после этого нам придется уже перемножать *столбцы* со столбцами, что поможет избежать неправильного сопоставления перемножаемых элементов.

Транспонирование (т. е. замена строк матрицы A ее столбцами) обозначается через $\tilde{A}$. В нашем примере

$$\tilde{A} = \begin{bmatrix} 33 & -24 & -8 \\ 16 & -10 & -4 \\ 72 & -57 & -17 \end{bmatrix}.$$

Итак, мы получим A^2 , умножив $\tilde{A}$ на A по столбцам. Для того чтобы указать, что мы имеем в виду не обыкновенное умножение матриц, а перемножение по столбцам, мы будем пользоваться знаком $\circ$:

$$A^2 = \begin{bmatrix} 33 & -24 & -8 \\ 16 & -10 & -4 \\ 72 & -57 & -17 \end{bmatrix} \circ \begin{bmatrix} 33 & 16 & 72 \\ -24 & -10 & -57 \\ -8 & -4 & -17 \end{bmatrix} = \\ = \begin{bmatrix} 129 & 80 & 240 \\ -96 & -56 & -189 \\ -32 & -20 & -59 \end{bmatrix}.$$

Повторяем процесс еще раз и получаем A^3 :

$$A^3 = \begin{bmatrix} 129 & -96 & -32 \\ 80 & -56 & -20 \\ 240 & -189 & -59 \end{bmatrix} \circ \begin{bmatrix} 33 & 16 & 72 \\ -24 & -10 & -57 \\ -8 & -4 & -17 \end{bmatrix} = \\ = \begin{bmatrix} 417 & 304 & 648 \\ -312 & -220 & -507 \\ -104 & -76 & -161 \end{bmatrix}.$$

Теперь мы составим правую часть уравнения (6):

$$6A^2 - 11A + 6I = \\ = \begin{bmatrix} 6 \cdot 129 - 11 \cdot 33 + 6 & 6 \cdot 80 - 11 \cdot 16 & 6 \cdot 240 - 11 \cdot 72 \\ -6 \cdot 96 + 11 \cdot 24 & -6 \cdot 56 + 11 \cdot 10 + 6 & -6 \cdot 189 + 11 \cdot 57 \\ -6 \cdot 32 + 11 \cdot 8 & -6 \cdot 20 + 11 \cdot 4 & -6 \cdot 59 + 11 \cdot 17 + 6 \end{bmatrix} =$$

Полученные элементы совпадают с элементами матрицы A^3 , и таким образом, справедливость равенства Гамильтона — Кели в рассматриваемом случае установлена.

Переходим теперь к определению собственных векторов («главных осей») нашей матрицы. Это означает, что нужно решить систему однородных линейных уравнений

$$(A - \lambda I) x = 0. \quad (2-6.7)$$

Этот процесс должен быть проведен для каждого λ_i , ибо каждому λ_i соответствует свой собственный вектор. Сначала мы возьмем корень $\lambda_1 = 1$. Сообразуясь с формулами (4.1), вычитаем 1 из диагональных элементов матрицы и получаем следующую систему линейных уравнений:

$$\left. \begin{aligned} 32x_1 + 16x_2 + 72x_3 &= 0, \\ -24x_1 - 11x_2 - 57x_3 &= 0, \\ -8x_1 - 4x_2 - 18x_3 &= 0. \end{aligned} \right\} \quad (2-6.8)$$

Мы знаем заранее, что определитель этой линейной системы равен нулю, так как именно это условие привело к определению собственных

значений. Но равенство нулю определителя системы n однородных линейных уравнений имеет вполне определенный смысл. Оно означает, что n уравнений не независимы друг от друга. Легко видеть, что в рассматриваемом случае второе уравнение не является следствием первого, однако третье уравнение является следствием первых двух. *Опустим* поэтому третье уравнение системы (8) и найдем решение остающихся двух (в общем случае $n - 1$) уравнений

$$32x_1 + 16x_2 + 72x_3 = 0, \quad -24x_1 - 11x_2 - 57x_3 = 0. \quad (2-6.9)$$

Если эти два уравнения удовлетворяются, то третье удовлетворяется автоматически.

Но теперь возникает новое затруднение: мы должны определить три неизвестных только из двух уравнений. С другой стороны, из однородного характера проблемы собственных значений следует, как мы уже знаем, что длина собственного вектора остается неопределенной. Это оставляет неопределенным множитель α : если x_1, x_2, x_3 есть какое-либо решение нашей задачи, то $\alpha x_1, \alpha x_2, \alpha x_3$ (при произвольном α) также есть годное решение. Но тогда мы можем воспользоваться произвольностью множителя α для целесообразного нормирования вектора x . Мы можем, например, принять, что $x_3 = 1$. Тогда система уравнений (9) принимает вид

$$32x_1 + 16x_2 + 72 = 0, \quad -24x_1 - 11x_2 - 57 = 0$$

и ее можно записать в неоднородной форме

$$32x_1 + 16x_2 = -72, \quad -24x_1 - 11x_2 = 57. \quad (2-6.10)$$

В общем случае матрицы n -го порядка первоначальная система из n однородных уравнений с n неизвестными может быть, таким образом, заменена системой $n - 1$ неоднородных уравнений с $n - 1$ неизвестными.

Эти уравнения могут быть решены (за исключением случая, когда они несовместны или избыточны) либо с помощью определителей (если n не превышает 3 или 4), либо же путем обращения матрицы*), если n велико. В нашем простом примере решение непосредственно получается, если использовать известные формулы решения системы двух линейных уравнений:

$$\left. \begin{aligned} a_{11}x_1 + a_{12}x_2 &= b_1, \\ a_{21}x_1 + a_{22}x_2 &= b_2, \\ x_1 &= \frac{b_1a_{22} - b_2a_{12}}{a_{11}a_{22} - a_{21}a_{12}}, \\ x_2 &= \frac{a_{11}b_2 - a_{21}b_1}{a_{11}a_{22} - a_{21}a_{12}} \end{aligned} \right\} \quad (2-6.11)$$

*) См. гл. III, § 7.

(в предположении, что $a_{11}a_{22} - a_{21}a_{12} \neq 0$). Этот процесс (для всех трех значений $\lambda = \lambda_1, \lambda_2, \lambda_3$) представлен в следующей таблице:

$\lambda = 1$ $32x_1 + 16x_2 = -72$ $-24x_1 - 11x_2 = 57$	$\lambda = 2$ $31x_1 + 16x_2 = -72$ $-24x_1 - 12x_2 = 57$	$\lambda = 3$ $30x_1 + 16x_2 = -72$ $-24x_1 - 13x_2 = 57$
$x_1 = \frac{-120}{32} = -\frac{15}{4}$ $x_2 = \frac{96}{32} = 3$ $x_3 = 1$	$x_1 = \frac{-48}{12} = -4$ $x_2 = \frac{39}{12} = \frac{13}{4}$ $x_3 = 1$	$x_1 = \frac{24}{-6} = -4$ $x_2 = \frac{-18}{-6} = 3$ $x_3 = 1$

Теперь мы составим таблицу результатов. Мы получили три собственных значения $\lambda_1, \lambda_2, \lambda_3$ и три соответствующих им вектора u_1, u_2, u_3 . Выпишем компоненты этих векторов в три отдельных столбца. Пользуясь произвольностью длин, мы можем умножить каждый из этих векторов на произвольную постоянную. Следовательно, мы можем освободиться от дробей и, например, вектор $(-15/4, 3, 1)$ заменить вектором $(-15, 12, 4)$. Наша таблица результатов выглядит так:

$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
u_1	u_2	u_3
-15	-16	-4
12	13	3
4	4	1

(2-6.12)

Хотя в нашем простом случае решение задачи нахождения собственных векторов и собственных значений было получено очень легко, нетрудно представить себе, что в случае матриц более высокого порядка расчеты гораздо более трудны и громоздки. Поэтому нужно найти методы, с помощью которых собственные значения и собственные векторы таких матриц можно было бы практически подсчитать с меньшей затратой труда. Такие методы мы позже рассмотрим весьма подробно.

Можно было бы думать, что наш анализ собственных значений теперь полностью закончен. Однако это не так. Существует еще одна черта матричной алгебры, которую мы до сих пор не рассматривали и которая имеет первостепенную важность. Заданная матрица A автоматически ассоциируется с другой матрицей, нераздельно связанной

с ней. Это — «транспонированная матрица», которую мы уже условились обозначать через $\tilde{A}$.

Расположение матричных элементов в строки и столбцы является в известной степени произвольным. Мы имеем перед собой квадрат, но что является «строкой» и что является «столбцом» зависит от того, как мы смотрим на него. Квадрат представляет полную симметрию в отношении расположения право — лево и верх — низ.

Вот мы смотрим на наш квадрат, расположенный обычным образом. Мы считаем, что одни элементы образуют «строки», другие — «столбцы» (рис. 4). Но, повернув квадрат на 90° и снова взглянув

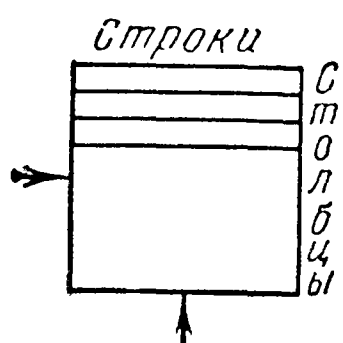


Рис. 4.

на него, мы найдем, что прежние строки стали столбцами, а столбцы — строками. Таким образом, при рассмотрении матрицы возникает некоторая двойственность. Матрица A неизбежно ассоциируется со своей транспонированной матрицей $\tilde{A}$, и мы не можем оперировать с матрицей A , не оперируя также одновременно с транспонированной матрицей $\tilde{A}$. В обыкновенной алгебре мы имеем несколько сходное явление

в поле комплексных чисел. Комплексное число $a + bi$ неизбежно ассоциируется с «комплексно сопряженным» числом $a - bi$, и оба они появляются одновременно во многих алгебраических проблемах. Например, при решении алгебраического уравнения с вещественными коэффициентами корень $a + bi$ всегда появляется вместе с корнем $a - bi$. Операция замены матрицы A на матрицу $\tilde{A}$ соответствует в некотором смысле замене в обыкновенной алгебре числа i на число $-i$.

Наш анализ собственных значений, таким образом, будет не полон, если не распространить его на транспонированную матрицу $\tilde{A}$. Здесь определяющим уравнением для задачи собственных значений становится уравнение

$$\tilde{A}x = \lambda x. \quad (2-6.13)$$

На первый взгляд эта задача совершенно отлична от предыдущей. Оказывается, однако, что решения обеих задач совпадают, поскольку речь идет о собственных значениях. Как мы видели, собственные значения λ матрицы удовлетворяют детерминантному уравнению (5.6). Но определитель не изменяет своего значения, если строки и столбцы меняются местами. Поэтому характеристический полином, соответствующий матрице $\tilde{A}$, в точности тот же, что и полином, соответствующий матрице A . Следовательно, собственные значения для A те же, что и собственные значения матрицы $\tilde{A}$. Но каждое собственное значение имеет теперь двойной аспект, поскольку мы будем пользоваться одним и тем же λ для решения уравнения $Ax = \lambda x$ и для решения уравнения $\tilde{A}x = \lambda x$. В первом случае мы найдем главные оси матрицы A , во втором — главные оси матрицы $\tilde{A}$.

Поэтому мы еще раз повторим все наши предыдущие операции, но используем теперь транспонированную матрицу $\tilde{A}$ вместо A :

$$\tilde{A} = \begin{bmatrix} 33 & -24 & 8 \\ 16 & -10 & 4 \\ 72 & -57 & 17 \end{bmatrix}.$$

Нет нужды повторять решение характеристического уравнения, так как оно остается тем же. Опять мы получаем три характеристических значения $\lambda = 1, 2, 3$. Однако таблица уравнений, приводящих к определению главных осей, теперь будет выглядеть следующим образом:

$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
$32x_1 - 24x_2 = 8$ $16x_1 - 11x_2 = 4$	$31x_1 - 24x_2 = 8$ $16x_1 - 12x_2 = 4$	$30x_1 - 24x_2 = 8$ $16x_1 - 13x_2 = 4$
$x_1 = \frac{8}{32} = \frac{1}{4}$ $x_2 = \frac{0}{32} = 0$ $x_3 = 1$	$x_1 = \frac{0}{12} = 0$ $x_2 = \frac{-4}{12} = -\frac{1}{3}$ $x_3 = 1$	$x_1 = \frac{-8}{-6} = \frac{4}{3}$ $x_2 = \frac{-8}{-6} = \frac{4}{3}$ $x_3 = 1$

Мы снова представим наши результаты в таблице, обозначив в ней собственные векторы транспонированной матрицы $\tilde{A}$ через v_1, v_2, v_3 . Мы опять можем избежать дробных чисел, произведя умножение на соответствующий множитель, ибо длина главных векторов остается неопределенной:

$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
v_1	v_2	v_3
1	0	4
0	-1	4
4	3	3

(2-6.14)

Таблицы (12) и (14), вместе взятые, дают полный анализ собственных значений нашей задачи. В общем случае мы имеем n скаляров, а именно собственные значения

$$\lambda = \lambda_1, \lambda_2, \dots, \lambda_n,$$

и $2n$ векторов, а именно n собственных векторов (главных осей) матрицы A :

$$u_1, u_2, \dots, u_n$$

и n собственных векторов (главных осей) матрицы $\tilde{A}$:

$$v_1, v_2, \dots, v_n.$$

В нашем численном примере полный анализ собственных значений заданной матрицы A может быть представлен в следующей таблице:

Полный анализ собственных значений

$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	
u_1	u_2	u_3	v_1	v_2	v_3	
— 15	— 16	— 4	1	0	4	(2-6.15)
12	13	3	0	— 1	4	
4	4	1	4	3	3	

Эти два ряда векторов находятся в замечательном взаимном отношении друг с другом. Векторы u сами между собой не обнаруживают какого-либо особенного внутреннего соотношения, это же относится и к векторам v , взятым сами по себе. Но если составим *скалярное произведение* *) одного из векторов u и одного из векторов v , например перемножим вектор u_1 с каждым из векторов v , то мы получим

$$\begin{aligned} u_1 v_1 &= -15 \cdot 1 + 12 \cdot 0 + 4 \cdot 4 = 1, \\ u_1 v_2 &= -15 \cdot 0 - 12 \cdot 1 + 4 \cdot 3 = 0, \\ u_1 v_3 &= -15 \cdot 4 + 12 \cdot 4 + 4 \cdot 3 = 0. \end{aligned}$$

Аналогично, сочетая u_2 со всеми векторами v , найдём

$$\begin{aligned} u_2 v_1 &= -16 \cdot 1 + 13 \cdot 0 + 4 \cdot 4 = 0, \\ u_2 v_2 &= -16 \cdot 0 - 13 \cdot 1 + 4 \cdot 3 = -1, \\ u_2 v_3 &= -16 \cdot 4 + 13 \cdot 4 + 4 \cdot 3 = 0. \end{aligned}$$

Наконец, сочетая u_3 со всеми векторами v , получаем

$$\begin{aligned} u_3 v_1 &= -4 \cdot 1 + 3 \cdot 0 + 1 \cdot 4 = 0, \\ u_3 v_2 &= -4 \cdot 0 - 3 \cdot 1 + 1 \cdot 3 = 0, \\ u_3 v_3 &= -4 \cdot 4 + 3 \cdot 4 + 1 \cdot 3 = -1. \end{aligned}$$

Если скалярное произведение двух векторов равно нулю, то это означает, на языке геометрии трехмерного пространства, что эти

*) Скалярным произведением вектора $u(x_1 \dots x_n)$ и вектора $v(y_1 y_2 \dots y_n)$ называют число (его обозначают uv), вычисляемое по формуле

$$uv = x_1 y_1 + x_2 y_2 + \dots + x_n y_n.$$

векторы *перпендикулярны* или *ортогональны* друг к другу. Мы видим, что *каждый вектор из ряда u ортогонален любому вектору из ряда v , за исключением одноименного с ним вектора*

$$u_i v_k = 0 \quad (i \neq k). \quad (2-6.16)$$

Скалярное произведение $u_1 v_1$ получилось равным 1, произведение $u_2 v_2$ равным -1 и произведение $u_3 v_3$ равным -1 . Легко видеть, что эти скалярные произведения можно было бы сделать равными любым числам, так как длины главных осей остаются неопределенными. Мы получим естественное нормирование длин главных осей, если потребуем, чтобы скалярные произведения $u_i v_i$ все были равны 1:

$$u_i v_i = 1. \quad (2-6.17)$$

Это требование все еще оставляет длины u_i неопределенными, но длины векторов v_i всегда могут быть подобраны так, чтобы удовлетворялось условие (17) (если оставить в стороне особый случай «неполноосных матриц», ср. § 11, когда некоторые произведения $u_i v_i$ оказываются равными нулю). Если первоначально некоторое $u_i v_i$ дает число c_i , то мы заменяем v_i на

$$\bar{v}_i = \frac{1}{c_i} v_i, \quad (2-6.18)$$

и тогда в новой системе векторов $\bar{v}_i$ уже будут выполняться условия (17).

В нашем численном примере вектор v_1 уже надлежаще нормирован, ибо $u_1 v_1$ случайно оказалось равным 1. Векторы v_2 и v_3 следует разделить на -1 , и таким образом, окончательная схема векторов v будет

v_1	v_2	v_3
1	0	—4
0	1	—4
4	—3	—3

(2-6.19)

Два ряда векторов, в которых каждый вектор одного ряда ортогонален всем векторам другого ряда, за исключением одноименного с ним, называются «биортогональными». Если, кроме того, скалярное произведение каждого вектора со своим одноименным вектором нормировано до 1, то мы говорим о биортогональных и нормированных рядах векторов.

Теперь мы опустим в таблице (15) вертикальные разделяющие линии, которые отделяют векторы u_i друг от друга и векторы v_i друг от друга. Тогда мы получаем (для векторов u) n столбцов элементов по n элементов в каждом столбце, которые вместе образуют квадратную матрицу из n^2 элементов. Другая квадратная матрица таблицы (15) состоит из координат векторов v_i . Мы получаем, таким образом, одну

матрицу U , включающую все векторы u_i , и одну матрицу V , включающую все векторы v_i . Следовательно, результаты полного анализа собственных значений могут быть записаны еще и другим образом: заданием n собственных значений $\lambda_1, \lambda_2, \dots, \lambda_n$ и двумя матрицами U и V по n строк и n столбцов в каждой. В нашем примере:

$$\lambda = 1, 2, 3,$$

$$U = \begin{bmatrix} -15 & -16 & -4 \\ 12 & 13 & 3 \\ 4 & 4 & 1 \end{bmatrix}, \quad V = \begin{bmatrix} 1 & 0 & -4 \\ 0 & 1 & -4 \\ 4 & -3 & -3 \end{bmatrix}. \quad (2-6.20)$$

Внутренние соотношения между этими двумя матрицами могут быть выражены в форме одного матричного уравнения

$$\tilde{U}V = I. \quad (2-6.21)$$

Действительно, перемножить две матрицы — это значит образовать скалярные произведения строк и столбцов. Строки первой матрицы умножаются на столбцы второй матрицы. Но транспонирование матрицы U приводит к тому, что *строки* матрицы $\tilde{U}$ фактически являются *столбцами* матрицы U . Таким образом, в левой стороне равенства (21) мы имеем произведения столбцов матрицы U на столбцы матрицы V . Тот факт, что эти произведения равняются 1 для соответствующих пар и 0 для всех остальных, означает, что диагональные элементы матрицы произведения равны 1, а все остальные элементы равны 0. Но это как раз есть определение единичной матрицы I . Легко понять, что содержание формулы (21) может быть записано также следующим образом:

$$\tilde{V}U = I. \quad (2-6.22)$$

7. Алгебраическое доказательство ортогональности собственных векторов

То, что мы наглядно показали только на численном примере, может быть подтверждено самым общим образом с помощью операций матричной алгебры. С этой целью мы сначала по-новому сформулируем общее содержание проблемы собственных значений:

$$Ax = \lambda x. \quad (2-7.1)$$

Это уравнение определяет каждый раз только *одну* главную ось, так как здесь λ представляет только одно из собственных значений, и для каждого $\lambda = \lambda_i$ уравнение должно быть отдельно решено. Уравнение (1), следовательно, должно быть использовано n раз для $\lambda = \lambda_1, \lambda_2, \dots, \lambda_n$.

Мы теперь объединим *все* собственные значения матрицы A в единой матричной записи. С этой целью введем матрицу Λ , определяемую

следующим образом:

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & & \lambda_n \end{bmatrix}. \quad (2-7.2)$$

Матрица такого вида, имеющая только диагональные элементы (все другие ее элементы равны нулю), называется «диагональной матрицей». Операции с такой матрицей особенно просты. Возьмем Λ в качестве первого сомножителя («умножение слева»). В силу общего правила умножения матриц мы получаем

$$\Lambda A = \begin{bmatrix} \lambda_1 a_{11} & \lambda_1 a_{12} & \dots & \lambda_1 a_{1n} \\ \lambda_2 a_{21} & \lambda_2 a_{22} & \dots & \lambda_2 a_{2n} \\ \dots & \dots & \dots & \dots \\ \lambda_n a_{n1} & \lambda_n a_{n2} & \dots & \lambda_n a_{nn} \end{bmatrix}. \quad (2-7.3)$$

Как видим, умножение слева на матрицу Λ состоит в том, что последовательные *строки* матрицы A умножаются соответственно на $\lambda_1, \lambda_2, \dots, \lambda_n$.

Возьмем теперь Λ в качестве второго сомножителя («умножение справа»):

$$A\Lambda = \begin{bmatrix} \lambda_1 a_{11} & \lambda_2 a_{12} & \dots & \lambda_n a_{1n} \\ \lambda_1 a_{21} & \lambda_2 a_{22} & \dots & \lambda_n a_{2n} \\ \dots & \dots & \dots & \dots \\ \lambda_1 a_{n1} & \lambda_2 a_{n2} & \dots & \lambda_n a_{nn} \end{bmatrix}. \quad (2-7.4)$$

Следовательно, умножение справа на матрицу Λ состоит в том, что последовательные *столбцы* матрицы A умножаются соответственно на $\lambda_1, \lambda_2, \dots, \lambda_n$.

Составим теперь матричное уравнение

$$AU = U\Lambda. \quad (2-7.5)$$

Если мы выпишем это уравнение поэлементно, то найдем, что первый столбец результирующей системы уравнений определяет первую главную ось u_1 (т. е. первый столбец матрицы U)*), второй столбец определяет вторую главную ось u_2 и т. д. Вся задача разыскания главных осей теперь выражена одним матричным уравнением, в котором U — искомая матрица.

*) Уравнение (5) представляет собой матричную запись системы n^2 уравнений — они получатся, если мы приравняем каждый элемент матрицы AU соответствующему элементу матрицы $U\Lambda$. Те n уравнений, которые возникают от сопоставления элементов, принадлежащих первым столбцам обеих матриц, представляют собой уравнения (4.2), в которых λ заменено на λ_1 , т. е. уравнения, определяющие первую главную ось.

Задача определения главных осей транспонированной матрицы $\tilde{A}$ подобным же образом выражается матричным уравнением

$$\tilde{A}V = V\Lambda, \quad (2-7.6)$$

в котором V — искомая матрица. Оно определяет все «сопряженные» оси $v_1, v_2, \dots, v_n$, которые являются последовательными столбцами матрицы V . Примем во внимание теперь, что из определения матричного умножения легко может быть выведено следующее фундаментальное правило матричной алгебры:

$$\widetilde{AB} = \tilde{B}\tilde{A}, \quad (2-7.7)$$

т. е. «транспонированная матрица произведения равна произведению транспонированных матриц сомножителей, взятых в обратном порядке». Такое же правило справедливо для любого числа сомножителей, например ¹⁾

$$\widetilde{(ABC)} = \tilde{C}\tilde{B}\tilde{A}. \quad (2-7.8)$$

Заметим, кроме того, что транспонирование диагональной матрицы Λ оставляет ее неизменной, так как на диагональные элементы не влияет процесс замены строк столбцами

$$\tilde{\Lambda} = \Lambda. \quad (2-7.9)$$

Учтем, наконец, и то, что «транспонирование транспонированной матрицы восстанавливает первоначальную матрицу»

$$\tilde{\tilde{A}} = A. \quad (2-7.10)$$

Это непосредственно следует из того факта, что замена строк на столбцы меняет последовательность индексов

$$\tilde{a}_{ik} = a_{ki},$$

а две таких замены восстанавливают первоначальную последовательность

$$\tilde{\tilde{a}}_{ik} = \tilde{a}_{ki} = a_{ik}.$$

Транспонируем теперь обе стороны равенства (6)

$$\tilde{V}A = \Lambda\tilde{V} \quad (2-7.11)$$

и помножим это равенство справа на U :

$$\tilde{V}AU = \Lambda\tilde{V}U. \quad (2-7.12)$$

¹⁾ Точно такое же правило имеет место и для «обращения», т. е. для возведения произведения матриц в степень -1 :

$$(ABC)^{-1} = C^{-1}B^{-1}A^{-1}.$$

С другой стороны, умножим слева равенство (5) на $\tilde{V}$:

$$\tilde{V}AU = \tilde{V}U\Lambda. \quad (2-7.13)$$

Так как левые части двух последних равенств равны, то

$$\Lambda \tilde{V}U = \tilde{V}U\Lambda, \quad (2-7.14)$$

или, обозначая произведение $\tilde{V}U$ через W :

$$\tilde{V}U = W, \quad (2-7.15)$$

получаем

$$\Lambda W = W\Lambda. \quad (2-7.16)$$

Это означает, что матрица W коммутативна с диагональной матрицей Λ . Мы можем также написать

$$\Lambda W - W\Lambda = 0, \quad (2-7.17)$$

что, в силу (3) и (4), означает

$$\begin{bmatrix} 0 & (\lambda_1 - \lambda_2) w_{12} & \dots & (\lambda_1 - \lambda_n) w_{1n} \\ (\lambda_2 - \lambda_1) w_{21} & 0 & \dots & (\lambda_2 - \lambda_n) w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ (\lambda_n - \lambda_1) w_{n1} & (\lambda_n - \lambda_2) w_{n2} & \dots & 0 \end{bmatrix} = 0. \quad (2-7.18)$$

Но тогда, предполагая, что корни λ_i характеристического уравнения все различны между собой, приходим к выводу, что все w_{ik} ($i \neq k$) должны равняться нулю, т. е. что W есть *диагональная* матрица:

$$W = \begin{bmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & w_n \end{bmatrix}. \quad (2-7.19)$$

Это уже доказывает биортогональность векторов u_i и v_i , так как (19) означает, что в силу определения (15) скалярное произведение любого u_i на v_k ($i \neq k$) равно нулю. Дополнительное условие, что все w_i равны 1, не может быть доказано, ибо оно зависит только от нормирования собственных векторов. Уравнения (5) и (6), определяющие матрицы U и V , таковы, что каждый столбец матриц U и V может быть умножен на произвольный множитель. Мы теперь распорядимся половиной этих множителей так, чтобы скалярные произведения $u_i v_i$ стали равны 1. Тогда W окажется единичной матрицей I , и мы получаем

$$\tilde{V}U = \tilde{U}V = I. \quad (2-7.20)$$

Мы имеем здесь пред собой основное соотношение между главными осями u_i и сопряженными осями v_i (т. е. главными осями транспонированной матрицы), выраженное в форме матричного равенства. Из

равенства (20) следует, что матрицы U и V находятся друг относительно друга в определенном взаимном отношении:

$$\begin{aligned} \tilde{V} &= U^{-1}, & V &= \tilde{U}^{-1}, \\ U &= \tilde{V}^{-1}, & \tilde{U} &= V^{-1}. \end{aligned} \quad (2-7.21)$$

(одна матрица есть «обратно транспонированная» к другой).

Еще одно замечательное соотношение можно получить, умножив (5) справа на $\tilde{V}$:

$$AU\tilde{V} = U\Lambda\tilde{V},$$

что дает, в силу (20) *),

$$A = U\Lambda\tilde{V}. \quad (2-7.22)$$

Аналогично

$$\tilde{A} = V\Lambda\tilde{U}. \quad (2-7.23)$$

Равенство (22) показывает, что исходную матрицу A можно получить, перемножив три матрицы, а именно: матрицу U , диагональную матрицу Λ и матрицу $\tilde{V}$.

Покажем теперь, что полное решение задачи собственных значений дает также решение задачи обращения матрицы. Запишем еще раз основное определяющее уравнение задачи собственных векторов для матрицы A :

$$Ax = \lambda x.$$

Умножая это равенство слева на A^{-1} , получаем

$$x = \lambda A^{-1}x \quad \text{или} \quad A^{-1}x = \lambda^{-1}x.$$

Это уравнение определяет собственные векторы матрицы A^{-1} , и мы приходим к заключению, что тот же вектор u_i , который был собственным вектором матрицы A , является также собственным вектором матрицы A^{-1} и что соответствующее собственное значение последней матрицы есть число, *обратное* первоначальному собственному значению λ_i . Следовательно, решение задачи собственных значений матрицы A^{-1} определяется следующими равенствами:

$$\begin{aligned} \lambda &= \frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \dots, \frac{1}{\lambda_n}, \\ U' &= U, \quad V' = V. \end{aligned} \quad (2-7.24)$$

Диагональная матрица Λ содержит теперь числа, *обратные* первоначальным числам λ , и является, следовательно, обратной к первоначальной матрице Λ . Если применим теперь соотношение (22) к матрице A^{-1} , то получим

$$A^{-1} = U\Lambda^{-1}\tilde{V}.$$

*) Если учесть, что из равенства $PQ = I$ следует равенство $QP = I$.

Таким образом, матрица, обратная матрице A , может быть выражена через собственные векторы и собственные значения матрицы A . Чтобы проиллюстрировать наши результаты на примере, обратимся к нашей прежней численной задаче; мы имеем

$$U = \begin{bmatrix} -15 & -16 & -4 \\ 12 & 13 & 3 \\ 4 & 4 & 1 \end{bmatrix}, \quad \Lambda = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}, \quad V = \begin{bmatrix} 1 & 0 & -4 \\ 0 & 1 & -4 \\ 4 & -3 & -3 \end{bmatrix}.$$

Следовательно,

$$\tilde{V} = \begin{bmatrix} 1 & 0 & 4 \\ 0 & 1 & -3 \\ -4 & -4 & -3 \end{bmatrix}, \quad \Lambda \tilde{V} = \begin{bmatrix} 1 & 0 & 4 \\ 0 & 2 & -6 \\ -12 & -12 & -9 \end{bmatrix}$$

и

$$\begin{aligned} U\Lambda\tilde{V} &= \begin{bmatrix} -15 & -16 & -4 \\ 12 & 13 & 3 \\ 4 & 4 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 4 \\ 0 & 2 & -6 \\ -12 & -12 & -9 \end{bmatrix} = \\ &= \begin{bmatrix} -15 & 12 & 4 \\ -16 & 13 & 4 \\ -4 & 3 & 1 \end{bmatrix} \circ \begin{bmatrix} 1 & 0 & 4 \\ 0 & 2 & -6 \\ -12 & -12 & -9 \end{bmatrix} = \begin{bmatrix} 33 & 16 & 72 \\ -24 & -10 & -57 \\ -8 & -4 & -17 \end{bmatrix}. \end{aligned}$$

Полученное произведение совпадает с заданной матрицей A .

Теперь мы повторяем нашу вычислительную схему, заменяя λ_i их обратными значениями:

$$\tilde{V} = \begin{bmatrix} 1 & 0 & 4 \\ 0 & 1 & -3 \\ -4 & -4 & -3 \end{bmatrix}, \quad \Lambda^{-1}\tilde{V} = \begin{bmatrix} 1 & 0 & 4 \\ 0 & \frac{1}{2} & -\frac{3}{2} \\ -\frac{4}{3} & -\frac{4}{3} & -1 \end{bmatrix},$$

$$U\Lambda^{-1}\tilde{V} = \begin{bmatrix} -15 & 12 & 4 \\ -16 & 13 & 4 \\ -4 & 3 & 1 \end{bmatrix} \circ \begin{bmatrix} 1 & 0 & 4 \\ 0 & \frac{1}{2} & -\frac{3}{2} \\ -\frac{4}{3} & -\frac{4}{3} & -1 \end{bmatrix} = \begin{bmatrix} -\frac{29}{3} & -\frac{8}{3} & -32 \\ 8 & \frac{5}{2} & \frac{51}{2} \\ \frac{8}{3} & \frac{2}{3} & 9 \end{bmatrix}.$$

Полученное произведение представляет собой матрицу A^{-1} . Для проверки составим произведение матрицы A на полученную матрицу A^{-1} :

$$AA^{-1} = \tilde{A} \circ A^{-1} = \begin{bmatrix} 33 & -24 & -8 \\ 16 & -10 & -4 \\ 72 & -57 & -17 \end{bmatrix} \circ \begin{bmatrix} -\frac{29}{3} & -\frac{8}{3} & -32 \\ 8 & \frac{5}{2} & \frac{51}{2} \\ \frac{8}{3} & \frac{2}{3} & 9 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Произведение этих двух матриц дает единичную матрицу I , это доказывает, что одна матрица есть обратная к другой.

Мы редко будем прибегать к этому методу нахождения обратной матрицы, так как обращение матрицы вообще требует гораздо меньше труда, чем решение полной проблемы собственных значений. Но если перед нами задача, при которой требуется глубокое исследование свойств заданной матрицы, то иногда необходимо найти все собственные векторы и собственные значения этой матрицы. В этом случае может оказаться ценным то, что задача обращения матрицы уже содержится в анализе собственных векторов и собственных значений. Кроме того, для критического исследования «почти особой системы уравнений», решение которой представляет большие практические трудности, связь проблемы собственных значений с проблемой обращения имеет неоценимое значение.

3. Геометрическая интерпретация проблемы собственных значений

Алгебраическая трактовка проблемы собственных значений проливает свет на некоторые характерные черты матричной алгебры. В обыкновенной алгебре произведения ab и ba взаимно заменяемы. В матричной алгебре это имеет место только при особых условиях. Мы обнаружили, например, что равенство произведений $W\Lambda$ и ΛW , где Λ есть диагональная матрица с различными диагональными элементами, возможно только в случае, когда W сама представляет собой диагональную матрицу. Мы обнаружили также, что равенство $AB=0$ не влечет за собой как необходимое следствие равенства нулю матрицы A или B . В нашем численном примере произведение

$$(A - I)(A - 2I)(A - 3I) = (A - I)(A^2 - 5A + 6I)$$

равнялось нулю. И все же ни $A - I$ ни $A^2 - 5A + 6I$ не равнялись нулю. Эти характерные отличия необходимо иметь в виду, когда мы оперируем с матрицами.

Чисто формальное оперирование с матрицами, давая эффектные результаты, легко может заслонить от нас более глубокие свойства матричной проблемы. Поэтому мы расширим картину и сделаем ее гораздо более содержательной, если свяжем с ней некоторую пространственную картину, которая дает *геометрическую* интерпретацию операций над матрицами. Операции с матрицами находятся в теснейшей связи с аналитической геометрией поверхностей второго порядка. Вся проблема собственных значений и собственных векторов имеет глубокую связь с геометрическими свойствами поверхностей второго порядка, которые в нашем обыкновенном трехмерном пространстве называются эллипсоидами и гиперболами.

Теория кривых и поверхностей второго порядка имеет давнюю и вдохновляющую историю. Греки проявили поразительную изобрета-

тельность в геометрическом исследовании конических сечений, которые они называли эллипсами, параболами и гиперболами. Аполлоний Пергамский, «великий геометр», владел всеми основными свойствами конических сечений. Он получил свои результаты частью аналитическими, частью проективными методами. Когда почти два тысячелетия спустя Декарт изобрел аналитическую геометрию и показал, как геометрические проблемы могут решаться при помощи алгебры, он создал новое и мощное орудие систематического исследования геометрических проблем. И все же и тогда не было известно ни одной основной теоремы из области конических сечений, которая не была бы открыта уже раньше греками с помощью одной только изобретательности, без помощи алгебры.

Греки не знали, какое практическое применение можно было бы найти для теории конических сечений. Прагматическая оценка вещей была им чужда, и они считали занятие геометрией привилегией интеллекта, черпающего наслаждение в самом себе, а не в каких-либо суежных выгодах. Но позднее это чисто эстетическое занятие геометрией принесло богатые плоды. Когда Кеплер дал новую оценку наблюдениям Тихо де Браге и пришел к заключению, что планеты обращаются вокруг Солнца не по окружностям, а по эллипсам, он опирался в своих вычислениях непосредственно на результаты Аполлония. Без греческой геометрии он не смог бы получить свои результаты. С другой стороны, теория тяготения Ньютона и теория движения планет не могли бы возникнуть без предшествующих результатов Кеплера. Мы видим, таким образом, прямую преемственность между исследованиями конических сечений у греков и теориями Кеплера, Ньютона, основаниями физики и техники XIX века и их развитием до современного состояния.

Однако астрономия и физика являются не единственными областями, в которых теория конических сечений играет выдающуюся роль. Аналитическая геометрия кривых и поверхностей второго порядка может быть распространена с пространств двух и трех измерений на пространства *любого* числа измерений. И тогда, как было обнаружено, вся теория линейных операторов — либо в виде теории обыкновенных линейных алгебраических уравнений, либо обыкновенных дифференциальных уравнений, или уравнений в частных производных, либо интегральных уравнений — может быть сформулирована в виде *геометрической* проблемы, относящейся к поверхностям второго порядка. Конические сечения были, так сказать, подняты с плоскости и помещены на гораздо более возвышенное основание. Пространство, с которым мы теперь оперируем, не является более пространством двух или трех измерений, а пространством многих измерений и может быть даже пространством бесконечно большого числа измерений. Но поверхности второго порядка, расположенные в этих многомерных пространствах, все же обладают теми же основными свойствами, которые греки открыли в своих исследованиях о конических сечениях.

В аналитической геометрии уравнение эллипса обыкновенно задается в следующем виде:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

Уравнение гиперболы обычно записывается в форме

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1.$$

Затем в аналитической геометрии трехмерного пространства прибавляется переменная z , и мы пишем уравнение эллипсоида в форме

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1 \text{ *)}.$$

Если мы желаем освободиться от случайных чисел 2 и 3 (числа измерений плоскости и пространства) и, исследуя пространства произвольного числа измерений, записывать наши уравнения таким образом, чтобы оставить свободным выбор *любого* числа измерений, то обычные обозначения аналитической геометрии должны быть основательно видоизменены. Мы не можем больше пользоваться буквами x , y , z для переменных, ибо нам не хватило бы букв. Кроме того, мы никогда не вскрыли бы внутренних законов, господствующих в наших уравнениях, если бы не ввели более систематизированные обозначения, отражающие однородную природу пространства в более адекватной форме. Обозначения x , y , z должны быть заменены на x_1 , x_2 , x_3 . Индексное обозначение имеет своим непосредственным следствием то, что переход к пространству любого числа измерений может быть выполнен с большой легкостью.

Воспользовавшись индексным обозначением координат, мы запишем уравнение эллипса в форме

$$\lambda_1 x_1^2 + \lambda_2 x_2^2 = 1, \quad (2-8.1)$$

а уравнение «эллипсоида» в форме

$$\lambda_1 x_1^2 + \lambda_2 x_2^2 + \lambda_3 x_3^2 = 1. \quad (2-8.2)$$

Обобщая эти уравнения на пространство n измерений, мы придем к следующему уравнению n -мерного «эллипсоида»:

$$\lambda_1 x_1^2 + \lambda_2 x_2^2 + \dots + \lambda_n x_n^2 = 1. \quad (2-8.3)$$

Тот факт что n заранее не задается численно, не является препятствием, — существенно лишь то, что мы знаем *закон*, по которому строится уравнение.

*) Существенно, что эти уравнения оказываются столь простыми по своей алгебраической структуре только потому, что мы выбираем специальную координатную систему — ее координатные оси являются осями симметрии изучаемых фигур.

Однако в физике и технике проблема изучения эллипса или эллипсоида встречается в иной форме. Если нашей целью является только изучение свойств эллипса, то естественно сразу же привести нашу систему координат в определенное соответствие с расположением этого эллипса — его большая и малая оси могут быть с самого начала взяты в качестве осей Ox и Oy нашей координатной системы. Однако обыкновенно дело обстоит не так. Мы имеем перед собой заданную систему координат, выбор которой определен другими соображениями, а исследуемый эллипс или эллипсоид может оказаться в наклонном положении относительно этой системы (рис. 5). Тогда

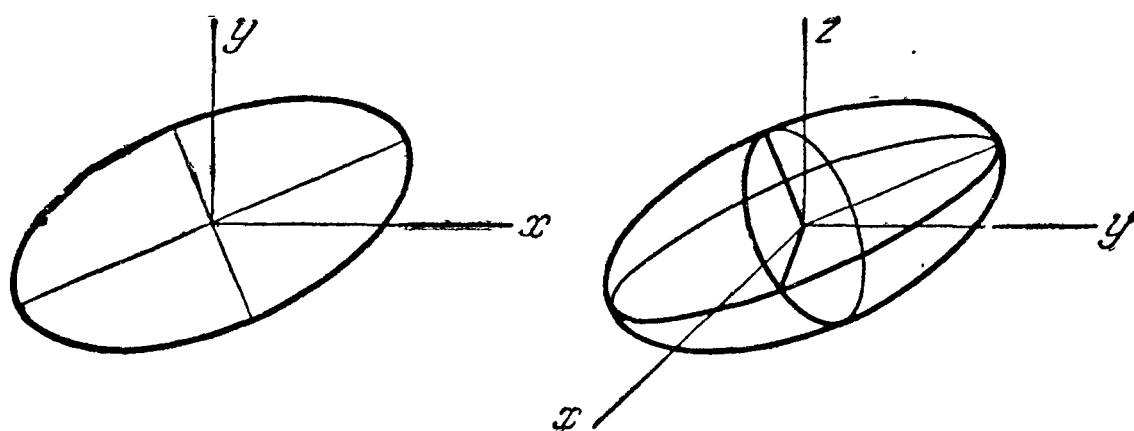


Рис. 5.

уравнение эллипса или эллипсоида записывается не так просто, как раньше. В случае эллипса или гиперболы появляется еще член xu , а в случае эллипсоида появляются произведения xu , uz и zx . Следовательно, уравнение эллипса имеет теперь три вместо двух членов, уравнение эллипсоида шесть вместо трех членов, а в общем n -мерном случае появляются произвольные произведения координат — это имеет своим следствием то, что уравнение содержит $\frac{1}{2}n(n+1)$ вместо n членов.

Нам нужно найти такое символическое обозначение всех членов уравнения n -мерного эллипсоида, чтобы оперирование с ними было более удобно. С этой целью мы воспользуемся вполне определенным методом. Хотя произведения xu и ux равны между собой (в силу коммутативного закона умножения), мы предпочитаем сохранить эти два члена раздельно. Следовательно, мы фактически усложняем положение, выписывая n^2 членов вместо $\frac{1}{2}n(n+1)$, однако теперь мы можем более эффективно проследить внутренний закон их образования. Уравнение эллипса будет записано следующим образом:

$$(a_{11}x_1 + a_{12}x_2)x_1 + (a_{21}x_1 + a_{22}x_2)x_2 = 1. \quad (2-8.4)$$

Уравнение эллипсоида примет вид

$$(a_{11}x_1 + a_{12}x_2 + a_{13}x_3)x_1 + (a_{21}x_1 + a_{22}x_2 + a_{23}x_3)x_2 + (a_{31}x_1 + a_{32}x_2 + a_{33}x_3)x_3 = 1. \quad (2-8.5)$$

В случае произвольного числа n измерений уравнение может быть записано следующим образом:

$$\begin{aligned} & (a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n)x_1 + \\ & + (a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n)x_2 + \\ & + \dots + (a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n)x_n = 1. \end{aligned} \quad (2-8.6)$$

Так как члены $x_i x_k$ и $x_k x_i$ можно объединить в один, то фактически в уравнении (6) имеет значение лишь сумма $a_{ik} + a_{ki}$, а так как она симметрична относительно индексов i и k , то целесообразно с самого начала принять, что

$$a_{ik} = a_{ki}. \quad (2-8.7)$$

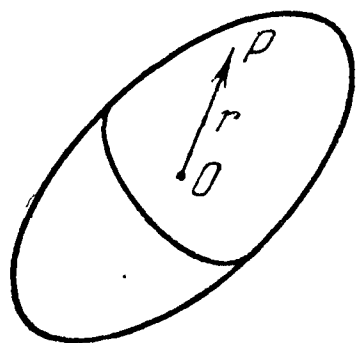
Матрица, которая нечувствительна к перемене порядка индексов, называется «симметрической матрицей». Например, матрица

$$A = \begin{bmatrix} 3 & 4 & -7 \\ 4 & -5 & 0 \\ -7 & 0 & 2 \end{bmatrix}$$

симметрична, так как $a_{12} = a_{21} = 4$, $a_{13} = a_{31} = -7$, $a_{23} = a_{32} = 0$. Транспонирование строк и столбцов в симметрической матрице ничего не меняет в ней; поэтому для нее справедливо равенство

$$\tilde{A} = A. \quad (2-8.8)$$

Такие матрицы обладают особенно важными свойствами, которые выделяют их из более широкого класса произвольных матриц.



Легко теперь видеть, что уравнение (6) общей центральной поверхности второго порядка может быть представлено в следующем матричном виде:

$$xAx = 1. \quad (2-8.9)$$

Рис. 6. Вектор $x = (x_1, x_2, x_3)$ (в случае трех измерений) и $x = (x_1, x_2, \dots, x_n)$ (в случае n измерений) называется «радиусом-вектором» r : он соединяет произвольную точку P поверхности с началом координат O (рис. 6).

Нашей задачей теперь будет определение главных осей (осей симметрии) поверхности второго порядка. Эта задача не возникает, если поверхность уже задана уравнением

$$\lambda_1 x_1^2 + \lambda_2 x_2^2 + \dots + \lambda_n x_n^2 = 1, \quad (2-8.10)$$

так как в этом случае главные оси совпадают с осями уже выбранной декартовой системы координат.

Что является специфической характеристикой главных осей поверхности второго порядка? В каждой точке поверхности может быть

проведена «нормаль» n , т. е. вектор, перпендикулярный к касательной плоскости поверхности в данной точке P . Векторы r и n вообще не параллельны друг другу. Только в отдельных исключительных направлениях (а именно в тех направлениях, которые обычно и выбирают в качестве координатных осей) нормали и радиусы-векторы оказываются параллельными друг другу (рис. 7). Отсюда и название «главные оси» для этой особой совокупности направлений.

Но, как известно, направляющие косинусы нормали n пропорциональны Ax^*). Таким образом, параллельность радиуса-вектора x и нормали Ax выражается равенством

$$Ax = \lambda x. \quad (2-8.11)$$

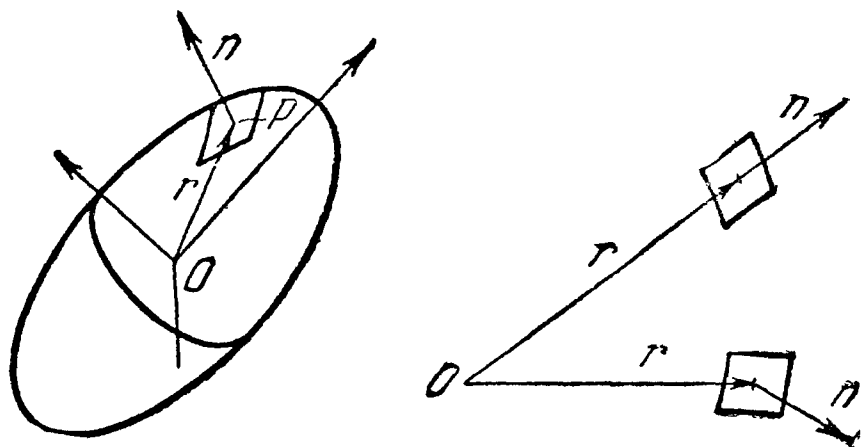


Рис. 7.

Это уравнение было основным уравнением теории собственных значений. Здесь мы к нему пришли, разыскивая главные оси поверхности второго порядка.

В прежнем нашем исследовании решение проблемы собственных значений состояло в следующем: находились n значений λ , при которых уравнение (11) имеет решение, затем мы получаем n соответствующих векторов $u_1, u_2, \dots, u_n$, при которых это уравнение удовлетворяется. Собственные значения $\lambda_1, \lambda_2, \dots, \lambda_n$ имеют следующее геометрическое значение. Найдем точку P с радиусом-вектором x , в которой главная ось пересекает поверхность второго порядка. Согласно уравнению поверхности

$$xAx = 1, \quad (2-8.12)$$

а согласно уравнению главных осей

$$Ax = \lambda x. \quad (2-8.13)$$

Умножая это уравнение на x^{**}), получаем

$$xAx = \lambda x^2 = 1, \quad (2-8.14)$$

что дает

$$x^2 = \frac{1}{\lambda}. \quad (2-8.15)$$

*) Изложение этого вопроса для пространства трех измерений читатель найдет, например, в книге М. Лагалли, Векторное исчисление, § 144.

**) То есть умножая обе части векторного равенства (13) скалярно на вектор x . Напомним, что скалярное произведение векторов обладает некоторыми (хотя и не всеми) свойствами произведения чисел; в частности, мы используем в дальнейшем свойство $(\lambda x) x = \lambda (xx)$. Число xx (скалярное произведение вектора x на самого себя) принято обозначать x^2 . Это число равно квадрату длины вектора x .

Выражение

$$x^2 = x_1^2 + x_2^2 + \dots + x_n^2 = r^2 \quad (2-8.16)$$

представляет собой квадрат расстояния от начала координат до точки P , в которой главная ось пересекает поверхность (рис. 8). Следовательно, λ_i есть величина, обратная квадрату расстояния этой точки от центра поверхности. Поэтому если какое-либо собственное

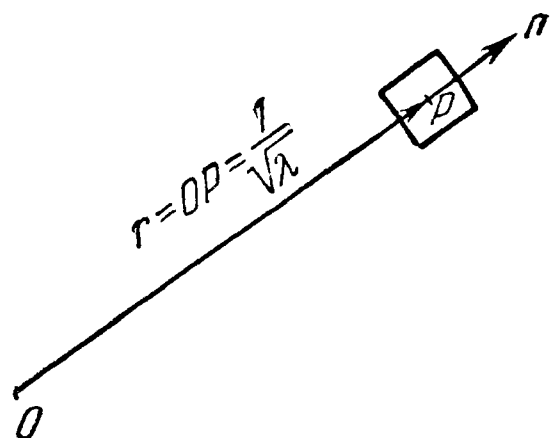


Рис. 8.

значение велико, то в направлении соответствующей главной оси поверхность второго порядка расположена близко к центру; если же собственное значение мало, то в направлении соответствующей главной оси поверхность далеко отстоит от своего центра.

Общая задача нахождения главных осей весьма упрощается в нашем случае в силу того факта, что A есть симметрическая матрица и, следовательно, $\tilde{A} = A$. В общем случае необходимо было бы отдельно находить главные оси для A и $\tilde{A}$. Но если $\tilde{A} = A$, то уравнение

$$Ax = \lambda x$$

разрешается одновременно с уравнением

$$\tilde{A}x = \lambda x.$$

Главные оси матриц A и $\tilde{A}$ теперь совпадают. Это имеет своим следствием то, что две матрицы U и V , которые характеризуют проблему главных осей в общем случае, оказываются равными:

$$V = U. \quad (2-8.17)$$

Но тогда основная связь между этими двумя матрицами

$$\tilde{U}V = 1$$

принимает вид

$$\tilde{U}U = 1. \quad (2-8.18)$$

Это равенство имеет следующий смысл. Составим скалярные произведения различных пар столбцов матрицы U . Скалярное произведение любого столбца на самого себя даст 1, а произведение какого-нибудь столбца на другой столбец даст 0. Это означает, что главные оси взаимно перпендикулярны, а длины собственных векторов нормированы до 1. Система n векторов n -мерного пространства, обладающих этим свойством, называется «ортонормальной системой».

Таким образом, мы убедились в том, что, «вообще говоря», поверхность второго порядка имеет n (и только n) главных осей, и эти оси между собой перпендикулярны. Мы можем поэтому принять

эти оси в качестве осей прямоугольной системы координат; таким образом мы приходим к пункту, с которого обычно начинается исследование в аналитической геометрии, где заранее принимается, что главные оси поверхности второго порядка совпадают с осями прямоугольной системы координат.

Собственно говоря, мы все же еще не доказали, что главные оси *вещественны*, так как собственные значения λ суть корни алгебраического уравнения n -й степени, и вообще говоря, эти корни могут быть комплексными числами. Однако имеет место следующее основное положение: *все собственные значения симметрической матрицы, составленной из вещественных чисел, вещественны*. Но если собственное значение λ вещественно, то соответствующее ему решение x уравнения

$$Ax = \lambda x \quad (2-8.19)$$

также должно быть вещественным.

То, что λ имеет вещественные значения, можно доказать следующим образом. Умножим равенство (19) на x^* (знак $*$ означает здесь переход к «комплексно сопряженному» количеству, т. е. замену i на $-i$):

$$x^* Ax = \lambda x^* x. \quad (2-8.20)$$

Так как алгебраическое равенство, содержащее комплексные числа, остается справедливым и после замены в нем каждого числа $\alpha + i\beta$ сопряженным ему числом $\alpha - i\beta$, то, выполнив эту операцию в равенстве (20), получим

$$xA^*x^* = \lambda^*xx^*. \quad (2-8.21)$$

Теперь мы воспользуемся следующим основным законом для транспонированных матриц:

$$xAy = y\tilde{A}x. \quad (2-8.22)$$

Применение его к левой части равенства (21) дает

$$x^*\tilde{A}^*x = \lambda^*xx^*. \quad (2-8.23)$$

Но если A — вещественная матрица, то $A^* = A$, ибо мнимая единица не содержится в элементах матрицы A , и следовательно, замена i на $-i$ оставляет эту матрицу неизменной. Кроме того, $\tilde{A} = A$ в силу симметрии матрицы. Следовательно,

$$x^*Ax = \lambda^*xx^* \quad (2-8.24)$$

и, вычитая (24) из (20), получаем (ввиду равенства левых частей)

$$(\lambda - \lambda^*)x^*x = 0. \quad (2-8.25)$$

Второй множитель не может равняться нулю, так как он является

суммой положительных чисел

$$x^*x = x_1^*x_1 + x_2^*x_2 + \dots + x_n^*x_n = |x_1|^2 + |x_2|^2 + \dots + |x_n|^2.$$

Мы получаем, таким образом, что

$$\lambda - \lambda^* = 0, \quad (2-8.26)$$

или

$$\lambda = \lambda^*. \quad (2-8.27)$$

Отсюда и следует, что λ есть *вещественное* число.

Мы убедились таким образом, что главные оси поверхности второго порядка вещественны. Важно заметить, что доказательство вещественности λ удалось провести благодаря условию

$$\tilde{A}^* = A. \quad (2-8.28)$$

В некоторых случаях приходится решать задачу нахождения главных осей, соответствующих матрице с *комплексными* элементами. Если эта матрица удовлетворяет условию (28), т. е. если транспонирование строк и столбцов с *одновременной* заменой i на $-i$ приводит опять к первоначальной матрице, то мы называем такую матрицу «эрмитовой» (или «самосопряженной»). Например, матрица

$$A = \begin{bmatrix} 3 & 4 + 2i & -8 + 5i \\ 4 - 2i & -5 & -3i \\ -7 - 5i & 3i & 2 \end{bmatrix}$$

есть эрмитова матрица.

Действительно, хотя транспонирование и не оставляет эту матрицу неизменной, но *транспонирование и одновременная замена i на $-i$ дает первоначальную матрицу*. Хотя собственные векторы такой матрицы будут уже не вещественными, а комплексными векторами, собственные значения все же вещественны, и те особые обстоятельства, которые имеют место в случае вещественных симметрических матриц, переносятся в область эрмитовых матриц. Эти матрицы играют в комплексной области ту же роль, что и симметрические матрицы в вещественной области. Если матрица имеет комплексные элементы, то ее симметрия не создает больших преимуществ, хотя остается верным, что матрица V совпадает при этом с U . Если же A эрмитова, то соотношение между V и U будет иметь вид

$$V = U^* \quad (2-8.29)$$

и, следовательно,

$$\tilde{U}U^* = I. \quad (2-8.30)$$

Это равенство вместе с вещественностью собственных значений обуславливает справедливость для таких матриц исключительных свойств ортонормальной системы главных осей.

9. Преобразование матрицы к главным осям

Геометрический подход к проблеме главных осей подчеркивает ту сторону теории, которая в чисто алгебраической трактовке не обнаруживается столь заметно. Когда мы рассматриваем матрицу в связи с поверхностью второго порядка, мы сразу видим, каков геометрический смысл главных осей. Но, кроме того, мы непосредственно усматриваем возможность изменения нашей системы координат. Так как главные оси взаимно ортогональны и имеют длину 1, то они представляются нам как естественная система координат, соответствующая заданной матрице. Поэтому целесообразно изучить теперь вопрос о *преобразовании координат*.

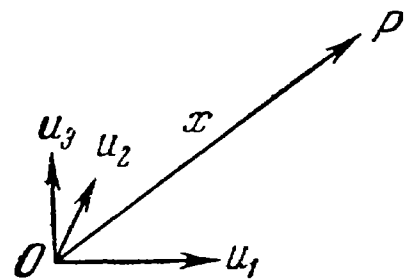


Рис. 9.

Исходные координаты вектора $x = OP$ в координатной системе, заданной векторами $u_1, u_2, \dots, u_n$ (рис. 9), суть

$$(x_1, x_2, \dots, x_n). \quad (2-9.1)$$

Заменив первоначально выбранные векторы u_i другой системой векторов $\bar{u}_1, \bar{u}_2, \dots, \bar{u}_n$ и представив радиус-вектор $x = OP$ в виде линейной комбинации этих новых векторов; $x = \bar{x}_1 \bar{u}_1 + \bar{x}_2 \bar{u}_2 + \dots + \bar{x}_n \bar{u}_n$, (рис. 10), получим новые координаты вектора x :

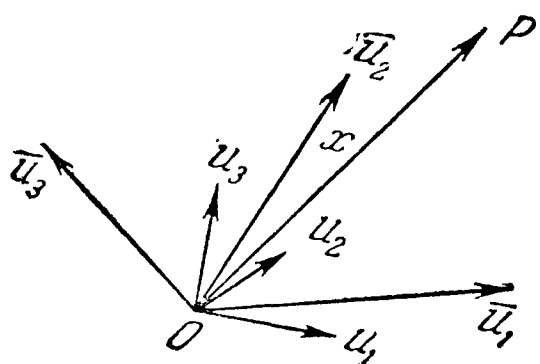


Рис. 10.

$$x = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n).$$

Общая задача преобразования координат представляется в следующей форме. Пусть дана совокупность «базисных векторов»

$$u_1, u_2, \dots, u_n, \quad (2-9.2)$$

не обязательно ортогональных между собой и не нормированных по длине. Следовательно, мы полагаем, что эти векторы u_i имеют произвольную длину и произвольное направление, удовлетворяя лишь *одному* условию, а именно: что они *линейно независимы*. Тогда любой вектор x может быть выражен линейной комбинацией этих векторов

$$x = x_1 u_1 + x_2 u_2 + \dots + x_n u_n \quad (2-9.3)$$

(при алгебраическом рассмотрении мы задавали этот вектор только его составляющими

$$x = (x_1, x_2, \dots, x_n) \quad (2-9.4)$$

и рассматривали его как совокупность величин x_i , которые характеризуются одним индексом).

Выбор базисных векторов $u_1, u_2, \dots, u_n$ более или менее случаен. С тем же основанием мы могли бы выбрать другую систему

выражения (7), то получим

$$\begin{aligned}
 x &= \bar{x}_1 (a_{11}u_1 + a_{21}u_2 + \dots + a_{n1}u_n) + \\
 &\quad + \bar{x}_2 (a_{12}u_1 + a_{22}u_2 + \dots + a_{n2}u_n) + \\
 &\quad + \dots + \bar{x}_n (a_{1n}u_1 + a_{2n}u_2 + \dots + a_{nn}u_n) = \\
 &= (a_{11}\bar{x}_1 + a_{12}\bar{x}_2 + \dots + a_{1n}\bar{x}_n) u_1 + \\
 &\quad + (a_{21}\bar{x}_1 + a_{22}\bar{x}_2 + \dots + a_{2n}\bar{x}_n) u_2 + \\
 &\quad + \dots + (a_{n1}\bar{x}_1 + a_{n2}\bar{x}_2 + \dots + a_{nn}\bar{x}_n) u_n.
 \end{aligned} \tag{2-9.10}$$

Первоначально вектор x был записан в виде

$$x = x_1 u_1 + x_2 u_2 + \dots + x_n u_n. \tag{2-9.11}$$

Новая форма (10) должна совпадать с (11), так как вектор не может иметь двух различных представлений в одной и той же системе векторов, если эти векторы линейно независимы *). Мы приходим поэтому к равенствам

$$\left. \begin{aligned}
 x_1 &= a_{11}\bar{x}_1 + a_{12}\bar{x}_2 + \dots + a_{1n}\bar{x}_n, \\
 x_2 &= a_{21}\bar{x}_1 + a_{22}\bar{x}_2 + \dots + a_{2n}\bar{x}_n, \\
 &\dots \dots \dots \checkmark \\
 x_n &= a_{n1}\bar{x}_1 + a_{n2}\bar{x}_2 + \dots + a_{nn}\bar{x}_n
 \end{aligned} \right\} \tag{2-9.12}$$

или, в матричном обозначении,

$$x = A\bar{x}^{**}). \tag{2-9.13}$$

Целесообразно слегка изменить наши обозначения. Так как первоначальные векторы $u_1, u_2, \dots, u_n$ не фигурируют в окончательной формуле (13), то мы можем опустить черточки над $\bar{u}_1, \bar{u}_2, \dots, \bar{u}_n$ и обозначить новые базисные векторы просто через $u_1, u_2, \dots, u_n$. Кроме того, составляющие этих векторов, в их разложении по векторам первоначальной системы координат, удобно обозначить через u_{ik} вместо a_{ik} . Поэтому матрицу преобразования целесообразно теперь обозначить через U , понимая это так, что последовательные столбцы этой матрицы представляют составляющие новых базисных векторов $u_1, u_2, \dots, u_n$. Формула (13), выражающая наше преобразование, теперь принимает вид

$$x = U\bar{x}. \tag{2-9.14}$$

*) Подробнее об этом сказано в § 10 этой главы.

**) Здесь x и $\bar{x}$ — только краткие обозначения двух столбцов: координат одного и того же вектора в двух различных координатных системах.

Если первоначальные базисные векторы нашей координатной системы представляют собой обычную прямоугольную систему координат, то для новых векторов $u_1, u_2, \dots, u_n$ это, вообще говоря, уже не будет иметь места. Однако может понадобиться сохранить прямоугольный характер и для новой системы координат. В этом случае мы должны наложить определенные условия на формулу преобразования (14). Новые векторы $u_1, u_2, \dots, u_n$ должны быть взаимно ортогональны, и их длины должны быть нормированы до 1. Это значит, что скалярное произведение любых двух различных векторов u должно равняться 0, а скалярное произведение любого вектора u на себя должно давать 1. Все эти условия содержатся в одном матричном равенстве

$$\tilde{U}U = I. \quad (2-9.15)$$

Это равенство совпадает с равенством (8.18), которое было получено раньше при решении задачи нахождения главных осей поверхности второго порядка. Эти оси образуют систему базисных векторов, которые перпендикулярны между собой и которые поэтому могут быть введены в качестве новой ортогональной координатной системы. Посмотрим, что случится с уравнением поверхности второго порядка в результате этого преобразования. Положим

$$x = U\bar{x} \quad (2-9.16)$$

и выполним эту подстановку в исходном уравнении (8.12), определяющем рассматриваемую поверхность второго порядка:

$$(U\bar{x})(AU\bar{x}) = 1. \quad (2-9.17)$$

Воспользуемся правилом перестановки (8.22):

$$xPy = y\tilde{P}x,$$

которое в применении к нашей задаче дает*)

$$\bar{x}(\tilde{A}\tilde{U})U\bar{x} = 1. \quad (2-9.18)$$

Так как, однако (ср. 7.7 и 8.8),

$$(\tilde{A}\tilde{U}) = \tilde{U}\tilde{A} = \tilde{U}A, \quad (2-9.19)$$

то мы получаем

$$\bar{x}\tilde{U}A\tilde{U}\bar{x} = 1. \quad (2-9.20)$$

Но, по определению главных осей,

$$AU = U\Lambda. \quad (2-9.21)$$

*) Если положить, что $AU = P$; $U\bar{x} = x$; $\bar{x} = y$.

Умножив слева на $\tilde{U}$, получаем

$$\tilde{U}AU = \tilde{U}UA \quad (2-9.22)$$

и, следовательно, *)

$$\bar{x}\Lambda\bar{x} = 1, \quad (2-9.23)$$

что означает

$$\lambda_1\bar{x}_1^2 + \lambda_2\bar{x}_2^2 + \dots + \lambda_n\bar{x}_n^2 = 1. \quad (2-9.24)$$

Таким образом, мы получаем традиционную форму уравнения поверхности второго порядка в том виде, как оно дается в аналитической геометрии в случае, когда заранее известно, что главные оси заданной поверхности второго порядка выбраны в качестве осей прямоугольной системы координат.

Тот факт, что в новой системе координат матрица A заменяется диагональной матрицей Λ , показывает, что преобразование к главным осям имеет глубокое влияние на A , преобразуя эту матрицу в особенно желательную форму, а именно в чисто *диагональную* форму. Оперирование с диагональными матрицами несравненно проще, чем оперирование с произвольными матрицами. Уравнения соответствующей системы разделяются по переменным и разрешаются непосредственно. С другой стороны, такое приведение матрицы к диагональной форме требует предварительного определения главных осей, что вообще нелегко выполнить. Тем не менее метод преобразования координат является исключительно важным средством матричного исследования. Если мы обладаем методом получения главных осей матрицы в каком-нибудь довольно *грубом приближении*, то хотя операция $\tilde{U}AU$ не превратит A в диагональную форму, но она *выделит по величине* диагональные элементы сравнительно с остальными элементами. Такая матрица представляет большие преимущества при численных расчетах, так как матричные задачи, соответствующие такой матрице, могут быть численно решены с помощью последовательности быстро сходящихся итераций.

Особого рассмотрения требует случай *кратных корней*. Собственные значения λ_i были получены при решении алгебраического уравнения n -й степени. Такое уравнение всегда имеет n корней, но может случиться, как мы уже сказали выше, что некоторые из этих корней совпадают. В этом случае мы говорим о «кратном корне», так как этот корень заменяет несколько различных корней. Случай совпадения корней можно представить себе как предельный, когда различные корни сближаются неограниченно между собой. Здесь снова геометрические представления помогут нам понять природу кратных корней. В уравнении эллипса равенство чисел λ_1 и λ_2 приводит к уравнению *окружности*

$$\lambda_1(x_1^2 + x_2^2) = 1.$$

*) В силу (15).

В случае эллипсоида равенство двух значений λ приводит к уравнению эллипсоида вращения

$$\lambda_1(x_1^2 + x_2^2) + \lambda_3 x_3^2 = 1,$$

а равенство всех трех значений λ — к уравнению сферы

$$\lambda_1(x_1^2 + x_2^2 + x_3^2) = 1.$$

Но если эллипс и превращается постепенно в окружность, то все же его главные оси не перестают существовать. Они превращаются в два взаимно перпендикулярных диаметра. Однако одна и та же окружность может служить предельным положением бесконечного числа эллипсов, и любая пара взаимно перпендикулярных диаметров окружности может служить главными осями. Подобным же образом в случае сферы любые три взаимно перпендикулярных диаметра могут служить главными осями сферы. То же имеет место и в случае многомерных пространств. Если в общем n -мерном случае m собственных значений сливаются в одно, то это означает, что в некотором m -мерном «подпространстве» проявляются «сферические условия». Любые m взаимно перпендикулярных осей могут быть выбраны в этом подпространстве в качестве главных осей поверхности второго порядка. Существование кратных корней не отменяет существование n различных взаимно перпендикулярных осей. Единственно, что имеет место, это то, что некоторые из этих осей уже не определены однозначно, а могут быть заменены другими равноправными им осями. Совпадение некоторых собственных значений в одно не связано с совпадением соответствующих осей. Взаимная перпендикулярность главных осей исключает возможность их слияния *).

При численных расчетах кратность корней всегда доставляет известные трудности. Если некоторые собственные значения, не сливаясь вместе, оказываются близкими друг к другу, то соответствующие им главные оси теоретически все еще однозначно определены. Но нахождение их с известной степенью точности становится все более трудным по мере уменьшения разности между двумя собственными значениями.

10. Косоугольная система координат

В наших предыдущих рассуждениях мы принимали, что поверхность второго порядка находится в наклонном положении относительно осей нашей координатной системы. Затем были найдены главные оси поверхности, и эти оси были приняты за новые оси координат. Преобразование к новым осям привело к превращению симметрической матрицы A в чисто диагональную матрицу Λ . Мы теперь сделаем

*) Приведенные здесь пояснения относятся только к случаю симметрической матрицы.

дальнейший шаг. До сих пор мы принимали, что наша исходная координатная система *прямоугольная* *), но ее оси не совпадают с главными осями поверхности второго порядка; это и вызвало необходимость преобразования координат. Так как и первоначальные координатные оси и главные оси поверхности образуют ортонормальные базисы, то произведенное преобразование было простым *вращением* системы осей, т. е. ортогональным преобразованием. Такое ортогональное преобразование характеризуется матричным уравнением

$$\tilde{U}U = I.$$

Может, однако, встретиться и более общее положение. Базисные векторы $u_1, u_2, \dots, u_n$ нашей первоначальной системы координат могут не быть ортогональны друг другу. Мы имеем тогда «косоугольную» систему осей. В этом случае поверхность второго порядка обязательно будет находиться в наклонном положении относительно координатных осей, так как главные оси этой поверхности взаимно ортогональны (рис. 11).

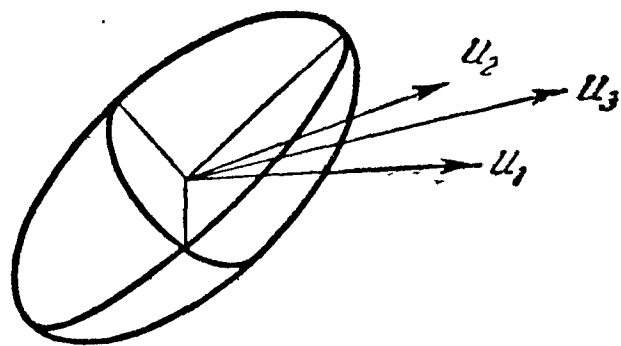


Рис. 11.

Так как в распространенных курсах аналитической геометрии используются почти исключительно прямоугольные системы осей **), то нашей первой задачей будет исследовать со всей общностью действия с косоугольными осями. Мы принимаем, что базисные векторы $u_1, u_2, \dots, u_n$ заданы как совокупность n линейно независимых векторов произвольной длины и произвольного направления. Это означает, что ни один из них не может быть получен, как линейная комбинация остальных векторов. Другими словами, *исключается возможность* существования линейного соотношения вида

$$\alpha_1 u_1 + \alpha_2 u_2 + \dots + \alpha_n u_n = 0, \quad (2-10.1)$$

в котором хотя бы один из коэффициентов α_i не равен нулю. Действительно, из такого соотношения следовало бы, что один из векторов, например u_n , если α_n отлично от нуля, может быть получен как линейная комбинация остальных векторов.

Из линейной независимости непосредственно следует, что произвольный вектор x , полученный как линейная комбинация базисных векторов

$$x = x_1 u_1 + x_2 u_2 + \dots + x_n u_n, \quad (2-10.2)$$

*) В частности, при записи уравнения поверхности второго порядка в форме (8.9).

**) В русской учебной литературе систематическое использование косоугольной координатной системы («общая аффинная координатная система») проводится, например, в следующих книгах: А. М. Лопшиц, Аналитическая геометрия, Учпедгиз, М., 1948; Б. Н. Делоне и Д. Н. Райков, Аналитическая геометрия, Гостехиздат, М., 1949.

не может иметь двух различных представлений. Ибо, если бы тот же вектор x мог быть представлен иным образом:

$$x = \xi_1 u_1 + \xi_2 u_2 + \dots + \xi_n u_n, \quad (2-10.3)$$

то, почленно вычитая равенства (2) и (3), мы получили бы

$$(x_1 - \xi_1) u_1 + (x_2 - \xi_2) u_2 + \dots + (x_n - \xi_n) u_n = 0. \quad (2-10.4)$$

Это, однако, устанавливало бы линейное соотношение вида (1), которое заранее было исключено. Представление (2) вектора x , таким образом, единственно.

Равенство (2) осуществляет некоторый *синтез*: из заданных базисных векторов $u_1, u_2, \dots, u_n$ мы получаем новый вектор путем умножения каждого из базисных векторов соответственно на $x_1, x_2, \dots, x_n$ и последующего векторного сложения полученных произведений. Но часто встречается *обратная* проблема. Задается вектор x и требуется найти, какая линейная комбинация базисных векторов образует заданный вектор. В этом случае мы говорим, что данный вектор *разлагается* в системе отсчета, заданной базисными векторами $u_1, u_2, \dots, u_n$. Здесь задается x и требуется найти коэффициенты $x_1, x_2, \dots, x_n$ линейной комбинации (2). Эти коэффициенты называются «координатами» вектора x в системе координат, заданной базисными векторами $u_1, u_2, \dots, u_n$.

С точки зрения такого разложения прямоугольные системы имеют огромное преимущество перед косоугольными системами. Если векторы $u_1, u_2, \dots, u_n$ образуют ортогональную и нормированную совокупность векторов, то они удовлетворяют равенствам

$$\left. \begin{aligned} u_i u_k &= 0, & (i \neq k), \\ u_i^2 &= 1 \end{aligned} \right\} \quad (2-10.5)$$

Эти равенства выражают в алгебраической форме тот геометрический факт, что любые два вектора этой совокупности взаимно-перпендикулярны, а длина каждого из них равна 1. В этой системе отсчета «скалярное» умножение вектора x на единичный вектор u_i дает

$$x u_i = (x_1 u_1 + x_2 u_2 + \dots + x_n u_n) u_i = x_i$$

и, следовательно,

$$x_i = x u_i. \quad (2-10.6)$$

Простое умножение вектора x на координатный вектор u_i дает скаляр x_i , который и является координатой вектора x в направлении базисного вектора u_i .

В косоугольной системе координат это обстоятельство уже не имеет места. Здесь теряется независимость, с которой действует

каждый отдельный базисный вектор ортогональной системы отсчета*), и нужно найти какие-нибудь другие средства для замены простого равенства (6). С этой целью мы строим новую совокупность векторов $v_1, v_2, \dots, v_n$, называемую «сопряженной» совокупностью**). Все, что мы получаем с помощью прямоугольных осей, может быть получено и при косоугольных осях, если только мы расширим заданную совокупность базисных векторов вдвое. Вместо оперирования с *одной* совокупностью векторов $u_1, u_2, \dots, u_n$, мы будем оперировать с *двойной* совокупностью векторов

$$\begin{pmatrix} u_1, u_2, \dots, u_n \\ v_1, v_2, \dots, v_n \end{pmatrix}. \quad (2-10.7)$$

Каждый вектор u_i ассоциируется с соответствующим вектором v_i , называемым «сопряженным» вектору u_i . Векторы v_i и u_i находятся в определенном двойственном отношении друг к другу, которое симметрично в обоих направлениях. Если даны векторы u_i , то их «сопряженными» будут векторы v_i . Если, наоборот, мы начинаем с векторов $v_1, v_2, \dots, v_n$ как базисных векторов, и строим сопряженные им, то мы получаем векторы $u_1, u_2, \dots, u_n$. Следовательно, совокупность, сопряженная сопряженной совокупности, совпадает с первоначальной.

Сопряженные векторы $v_1, v_2, \dots, v_n$ определяются однозначно. Они находятся в биортогональном отношении с первоначальными векторами $u_1, u_2, \dots, u_n$ — это означает, что любой вектор u_i одной совокупности и любой вектор v_k другой совокупности (исключая ему самому сопряженный вектор v_i) взаимно ортогональны. Кроме того, длина сопряженных векторов v_i нормирована условием, что скалярное произведение $u_i v_i$ каждого вектора со своим сопряженным равно 1:

$$\left. \begin{aligned} u_i v_k &= 0, & (i \neq k). \\ u_i v_i &= 1 \end{aligned} \right\} \quad (2-10.8)$$

Таким образом, преимущество прямоугольной совокупности векторов состоит в том, что исчезает необходимость построения второй совокупности сопряженных векторов, так как если векторы u_i сами удовлетворяют соотношениям ортогональности (5), то отсюда следует, что векторы v_i совпадают с векторами u_i :

$$v_i = u_i. \quad (2-10.9)$$

Поэтому прямоугольная и нормированная совокупность векторов называется «самосопряженной». Для всякой другой совокупности векторов совокупность v_i должна быть особо построена. Однако после того,

*) То есть исчезает возможность вычислить k -ю координату, используя только k -й координатный вектор.

**) Как это было указано в § 7 [ср. (7.6)].

как это построение произведено, мы можем работать с косоугольной системой отсчета так же легко и эффективно, как и с ортогональной системой.

Поэтому мы будем считать, что всякий раз, когда задается косоугольная система отсчета $u_1, u_2, \dots, u_n$, мы одновременно имеем ассоциированную с ней систему $v_1, v_2, \dots, v_n$ и будем всегда рассматривать обе векторные системы совместно, а не одну отдельно от другой:

$$\begin{pmatrix} u_1, u_2, \dots, u_n \\ v_1, v_2, \dots, v_n \end{pmatrix}; \quad u_i v_k = 0 \quad (i \neq k); \quad u_i v_i = 1. \quad (2-10.10)$$

Проблема разложения вектора x в системе отсчета базисных векторов u_i теперь получает простое решение. Мы умножаем равенство

$$x = x_1 u_1 + x_2 u_2 + \dots + x_n u_n$$

на вектор v_i и получаем, на основании соотношений биортогональности (8),

$$x v_i = x_i$$

и, следовательно,

$$x_i = x v_i. \quad (2-10.11)$$

Скалярное умножение вектора x на сопряженный вектор v_i дает число x_i , т. е. координату вектора x в направлении базисного вектора u_i .

Векторы u_i и v_i столь тесно связаны теперь друг с другом, что одна система векторов не имеет смысла без другой. Если мы пользуемся u_i как базисной системой для разложения вектора x , то с тем же основанием мы можем воспользоваться векторами v_i , т. е. разложением

$$x = \xi_1 v_1 + \xi_2 v_2 + \dots + \xi_n v_n. \quad (2-10.12)$$

Следовательно, в косоугольной системе координат мы имеем два представления вектора x : одно, обозначенное латинскими буквами, а другое, обозначенное греческими буквами. Если (12) умножить скалярно на u_i , то мы получим

$$\xi_i = x u_i.$$

Вектор x один и тот же в обоих случаях; изменилась лишь *система отсчета*, в которой произведено разложение. Мы имеем не два различных вектора, а два сопряженных представления одного и того же вектора, разложенного один раз по осям первоначальной, а другой раз по осям сопряженной системы:

$$x = x_1 u_1 + x_2 u_2 + \dots + x_n u_n = \xi_1 v_1 + \xi_2 v_2 + \dots + \xi_n v_n. \quad (2-10.13)$$

Однако при алгебраических операциях мы *опускаем* обозначения

базисных векторов $u_1, u_2, \dots, u_n$, оставляя только составляющие. Мы, следовательно, пишем

$$x = (x_1, x_2, \dots, x_n). \quad (2-10.14)$$

Подобным же образом, когда мы оперируем с сопряженной совокупностью осей, мы опускаем обозначения базисных векторов $v_1, v_2, \dots, v_n$ и пишем только составляющие $\xi_1, \xi_2, \dots, \xi_n$. Но эти составляющие отличны, конечно, от прежних составляющих, и таким образом, получающийся вектор естественно теперь обозначить не x , а ξ :

$$\xi = (\xi_1, \xi_2, \dots, \xi_n). \quad (2-10.15)$$

Хотя это обозначение вполне логично и непротиворечиво с алгебраической точки зрения, необходимо иметь в виду, что эти два «алгебраических вектора» x и ξ представляют собой только *два равноправных представления одного единого* геометрического образа, а именно вектора x , в двух взаимно связанных системах отсчета, с которыми мы оперируем.

С точки зрения алгебры вектор x целесообразно записывать как *пару* векторов (x, ξ) . В обычной прямоугольной системе координат это удвоение числа составляющих не возникает, так как вектор x и вектор ξ совпадают. Но в косоугольной системе вектор x непременно ассоциирован с двойственным ему вектором ξ .

Рассмотрим теперь, как выражается скалярное произведение двух векторов x и y через их координаты; пусть

$$\begin{aligned} x &= x_1 u_1 + x_2 u_2 + \dots + x_n u_n = \xi_1 v_1 + \xi_2 v_2 + \dots + \xi_n v_n, \\ y &= y_1 u_1 + y_2 u_2 + \dots + y_n u_n = \eta_1 v_1 + \eta_2 v_2 + \dots + \eta_n v_n. \end{aligned}$$

Учитывая соотношения (8), связывающие векторы u_i и v_i , целесообразно выражать один из перемножаемых векторов в координатной системе $u_1 \dots u_n$, а другой — в сопряженной с ней координатной системе $v_1, v_2, \dots, v_n$. С алгебраической точки зрения мы составляем, таким образом, произведения алгебраических векторов x и η , либо ξ и y (но не произведения xu или $\xi\eta$):

$$\begin{aligned} xy &= (x_1 u_1 + x_2 u_2 + \dots + x_n u_n) (\eta_1 v_1 + \eta_2 v_2 + \dots + \eta_n v_n) = \\ &= x_1 \eta_1 + x_2 \eta_2 + \dots + x_n \eta_n = x\eta, \\ xy &= (\xi_1 v_1 + \xi_2 v_2 + \dots + \xi_n v_n) (y_1 u_1 + y_2 u_2 + \dots + y_n u_n) = \\ &= \xi_1 y_1 + \xi_2 y_2 + \dots + \xi_n y_n = \xi y. \end{aligned}$$

Мы вычисляем двумя разными способами одно и то же скалярное произведение векторов x и y , но по-разному обозначаем полученный результат; таким образом,

$$x\eta = \xi y. \quad (2-10.16)$$

Обычное правило скалярного умножения двух векторов *), состоящее в умножении соответствующих составляющих и последующего суммирования частных произведений:

$$\begin{aligned} a &= (a_1, a_2, \dots, a_n), \\ b &= (b_1, b_2, \dots, b_n), \\ ab &= a_1 b_1 + a_2 b_2 + \dots + a_n b_n, \end{aligned}$$

сохраняет в косоугольной координатной системе свой внешний вид. Но при этом нужно помнить, что, составляя по этому правилу скалярное произведение двух векторов, следует использовать различные двойственные друг другу алгебраические представления перемножаемых векторов. Выражения

$$\begin{aligned} xy &= x_1 y_1 + x_2 y_2 + \dots + x_n y_n, \\ \xi \eta &= \xi_1 \eta_1 + \xi_2 \eta_2 + \dots + \xi_n \eta_n \end{aligned}$$

не имеют никакого смысла **). Один множитель должен быть разложен в системе u , а другой в системе v . (В тензорном исчислении координаты $x_1, x_2, \dots, x_n$ вектора x называются «ковариантными компонентами», а координаты $\xi_1, \xi_2, \dots, \xi_n$ этого же самого вектора называют «контравариантными компонентами» вектора x . Различие между этими двумя совокупностями составляющих отмечается в тензорном исчислении не с помощью латинского и греческого алфавитов, а записью индексов один раз сверху, а другой раз — снизу

$$\begin{aligned} x &= (x_1, x_2, \dots, x_n) \quad (\text{ковариантный вектор}), \\ x &= (x^1, x^2, \dots, x^n) \quad (\text{контравариантный вектор}). \end{aligned}$$

Скалярное произведение векторов x и y принимает вид

$$xy = x^1 y_1 + x^2 y_2 + \dots + x^n y_n = x_1 y^1 + x_2 y^2 + \dots + x_n y^n.$$

Это обозначение, однако, менее удобно при матричных операциях.)

11. Преобразование к главным осям в случае, когда поверхность задана в косоугольной системе

Предположим, что поверхность второго порядка задана в косоугольной системе координат, имеющей базисные векторы:

$$\begin{pmatrix} u_1, u_2, \dots, u_n \\ v_1, v_2, \dots, v_n \end{pmatrix}. \quad (2-11.1)$$

*) Заданных своими прямоугольными координатами.

**) Ибо значения этих выражений будут изменяться при переходе к другой системе координат, тогда как выражения (10.16), как это можно доказать, остаются неизменными при любом преобразовании координат (ср. далее §§ 12 и 13).

Радиус-вектор x может быть задан в двойкой форме:

$$\left. \begin{aligned} x &= (x_1, x_2, \dots, x_n), \\ \xi &= (\xi_1, \xi_2, \dots, \xi_n). \end{aligned} \right\} \quad (2-11.2)$$

Уравнение поверхности второго порядка в обыкновенных прямоугольных координатах имело вид

$$x \cdot Ax = 1. \quad (2-11.3)$$

В косоугольной системе координат скалярное произведение может быть образовано лишь в двойственном представлении, когда один вектор задается в базисной системе, а другой — в сопряженной ей системе. Следовательно, теперь мы имеем

$$\xi \cdot Ax = 1, \quad (2-11.4)$$

или, в подробной записи,

$$\begin{aligned} &\xi_1 (a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n) + \\ &+ \xi_2 (a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n) + \\ &+ \dots + \\ &+ \xi_n (a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n) = 1. \end{aligned} \quad (2-11.5)$$

Теперь в левую часть входят n^2 различных членов, они не могут быть приведены к $\frac{1}{2}n(n+1)$ членам, так как нельзя сделать приведения членов $a_{ik}\xi_i x_k$ и $a_{ki}\xi_k x_i$. Матрица A уже не симметрическая, а произвольная; она содержит n^2 различных элементов.

Как и раньше, главные оси поверхности второго порядка суть те направления, в которых радиус-вектор и нормаль к поверхности в точке ее пересечения с этим радиусом параллельны; направляющие косинусы нормали пропорциональны координатам вектора Ax , и эта пропорциональность нормали и радиуса-вектора выражается равенством

$$Ax = \lambda x. \quad (2-11.6)$$

Однако каждый вектор теперь имеет двойственный аспект, ибо может быть выражен как через первоначальные базисные векторы u_i , так и через сопряженные векторы v_i . Мы должны поэтому получить уравнения главных осей не только в первоначальной координатной системе, но и в сопряженной. Это значит, что нужно дополнительно составить уравнение, которое определит ξ вместо x . Учитывая алгебраическое тождество

$$xAy = y\tilde{A}x,$$

мы можем записать основное уравнение (4) в иной форме

$$x\tilde{A}\xi = 1. \quad (2-11.7)$$

черточек сверху, т. е. будем писать просто $u_1, u_2, \dots, u_n$. Эти векторы алгебраически характеризуются матрицей

$$U = \begin{array}{c} \begin{array}{cccc} u_1 & u_2 & \dots & u_n \end{array} \\ \left[\begin{array}{cccc} u_{11} & u_{12} & \dots & u_{1n} \\ u_{21} & u_{22} & \dots & u_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ u_{n1} & u_{n2} & \dots & u_{nn} \end{array} \right] \end{array} \quad (2-11.12)$$

Но эти же самые векторы, разложенные в сопряженной системе, алгебраически определяют другую совокупность алгебраических векторов — мы их обозначим $v_1, v_2, \dots, v_n$:

$$V = \begin{array}{c} \begin{array}{cccc} v_1 & v_2 & \dots & v_n \end{array} \\ \left[\begin{array}{cccc} v_{11} & v_{12} & \dots & v_{1n} \\ v_{21} & v_{22} & \dots & v_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ v_{n1} & v_{n2} & \dots & v_{nn} \end{array} \right] \end{array} \quad (2-11.13)$$

Столбцы матрицы U и столбцы матрицы V изображают одну и ту же совокупность n геометрических векторов. Тот факт, что любые две главные оси ортогональны одна к другой, означает, что скалярное произведение любого столбца матрицы U на какой-нибудь столбец матрицы V , за исключением ему соответствующего, равняется нулю. Из того, что длина векторов, определяющих главные оси нормирована до 1, следует, что произведение любого столбца матрицы U на соответствующий столбец матрицы V дает 1. В матричном выражении

$$\tilde{U}V = I. \quad (2-11.14)$$

Мы встречали это равенство раньше как результат алгебраических операций. Новое понимание, которое мы получили благодаря геометрической трактовке, состоит в том, что это равенство выражает ортогональность главных осей поверхности второго порядка. Раньше эта ортогональность была выражена в форме

$$\tilde{U}U = I,$$

и матрица A была симметричной. Новое равенство (14) получилось потому, что система базисных векторов, по которым мы разлагаем, уже не прямоугольна, а косоугольна.

Теперь мы введем главные оси поверхности второго порядка в качестве новой системы координат. Здесь уже не нужно рассматривать вместе с первоначальной и сопряженную ей систему, так как новая система координат «ортонормальна» (ортогональна и нормирована по

длине) и поэтому самосопряженна. Преобразование к новым осям задается равенствами

$$x = U\bar{x}, \quad \xi = V\bar{x}. \quad (2-11.15)$$

Вводя это преобразование в (4), получаем

$$V\bar{x} \cdot AU\bar{x} = 1 \quad (2-11.16)$$

и, в силу закона транспонирования,

$$\bar{x} (\tilde{A}U) V\bar{x} = \bar{x} \tilde{U} \tilde{A} V\bar{x} = 1. \quad (2-11.17)$$

Так как, однако, по определению матрицы V ,

$$\tilde{A}V = V\Lambda, \quad (2-11.18)$$

то умножение слева на $\tilde{U}$ дает

$$\tilde{U} \tilde{A} V = \tilde{U} V \Lambda, \quad (2-11.19)$$

и таким образом, согласно (14) и (17), уравнение поверхности второго порядка в новой системе координат принимает вид

$$\bar{x} \Lambda \bar{x} = 1. \quad (2-11.20)$$

Это уравнение совпадает с ранее полученным (9.23) при преобразовании поверхности второго порядка к ее главным осям. Прежде мы исходили из симметрической матрицы, тогда как теперь исходным пунктом была произвольная матрица. Конечный результат должен быть одним и тем же, ибо в обоих случаях мы рассматриваем одну и ту же поверхность второго порядка; при окончательном преобразовании, после введения главных осей самой поверхности в качестве координатных осей, все, что было связано с прежней совокупностью осей, исчезает и, следовательно, не может быть никакого различия в окончательном результате, независимо от того, начинали ли мы с прямоугольной или с косоугольной совокупности осей.

Заканчивая анализ, мы можем сказать, что проблема главных осей поверхностей второго порядка, рассмотренная в косоугольной системе отсчета, показывает, что даже несимметрические матрицы могут быть преобразованы к чисто диагональной форме*). Это преобразование имеет вид $\tilde{U} \tilde{A} V = \Lambda$, или также

$$\tilde{V} A U = \Lambda. \quad (2-11.21)$$

Существует, однако, фундаментальное различие между случаями косоугольной системы и прямоугольной. В прямоугольном случае ($\tilde{A} = A$) мы могли доказать, что все собственные значения и собственные векторы *вещественны*. Поэтому переход от одной системы координат

*) Если они удовлетворяют некоторым специальным требованиям.

к другой осуществляем действительным вращением в n -мерном пространстве. В общем случае это уже не так. Собственные значения λ_i являются, вообще говоря, *комплексными числами*, и следовательно, элементы матрицы Λ , вообще говоря, комплексны. Подобным же образом элементы матриц U и V представляют собой, вообще говоря, также комплексные числа.

Другое существенное отличие связано с вопросом о совпадении собственных векторов. Пока собственные значения различны, соответствующие собственные векторы также различны. Однако в случае кратных собственных значений возникают специфические явления. В случае симметрической матрицы совпадение собственных значений, как мы видели, *не* означает соответствующего совпадения собственных векторов. Оно означает только некоторую неопределенность в выборе собственных векторов, поскольку в определенном подпространстве m измерений *любые* m осей могут быть взяты в качестве собственных осей. Иначе дело обстоит в несимметрическом случае.

В качестве интересного примера рассмотрим три матрицы:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 1 & 2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & -1 & 3 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix}. \quad (2-11.22)$$

Для каждой из них «детерминантное уравнение», определяющее собственные значения λ_i , сводится к уравнению

$$(\lambda - 1)^3 = 0, \quad (2-11.23)$$

которое показывает, что $\lambda = 1$ есть единственное возможное собственное значение, кратности 3. Три, вообще говоря, различных собственных значения матрицы из 3×3 элементов в нашем случае сливаются в один.

Но в первом случае заданная матрица есть единичная матрица. Соответствующая поверхность второго порядка есть сфера. Здесь мы можем выбрать любые три взаимно ортогональные оси в качестве главных осей заданной матрицы. Три главные оси существуют и они совершенно различны, хотя их направление и не определено однозначно; бесконечное множество ортогональных систем может быть выбрано в качестве системы главных осей.

Рассмотрим теперь вторую матрицу. Уравнения для определения главных осей имеют теперь следующий вид:

$$\begin{aligned} x_1 + x_2 + 2x_3 &= x_1, \\ x_2 &= x_2, \\ x_3 &= x_3. \end{aligned}$$

Эта система сводится к единственному уравнению

$$x_2 + 2x_3 = 0,$$

которое имеет два независимых решения:

$$\begin{aligned} x_2 &= x_3 = 0, \\ x_2 &= -2, \quad x_3 = 1. \end{aligned}$$

В первом случае x_1 не может равняться нулю, ибо иначе весь вектор исчезал бы. Следовательно, мы можем принять

$$u_1 = (1, 0, 0)$$

в качестве нашего первого решения. Второе решение оставляет x_1 произвольным, и мы можем в качестве второго решения взять вектор

$$u_2 = (0, -2, 1).$$

Другая возможность

$$u_3 = (a, -2, 1)$$

при произвольном a приводит к вектору, который является лишь линейной комбинацией уже найденных решений u_1 к u_2 :

$$u_3 = au_1 + u_2.$$

Мы видим, следовательно, что в этом случае кратное собственное значение $\lambda = 1$ ассоциируется только с двумя линейно независимыми собственными векторами.

Теперь рассмотрим третью матрицу. Здесь уравнения для определения главных осей имеют вид

$$\begin{aligned} x_1 - x_2 + 3x_3 &= x_1, \\ x_2 + 2x_3 &= x_2, \\ x_3 &= x_3. \end{aligned}$$

Эта система эквивалентна системе

$$\begin{aligned} x_3 &= 0, \\ -x_2 + 3x_3 &= 0. \end{aligned}$$

Решением ее является вектор

$$u_1 = (1, 0, 0).$$

В этом случае, следовательно, число собственных векторов равно только единице. Три, вообще говоря, независимых собственных вектора слились в один.

Таким образом, слияние собственных значений может одновременно вызвать слияние соответствующих собственных векторов — явление, не имеющее аналога в области вещественных симметрических матриц. Возникает вопрос, можно ли было бы, хотя бы в частных случаях, предвидеть такое своеобразное поведение собственных векторов, не входя в подробное решение задачи собственных значений. Что это именно так, мы увидим, если обратим внимание на равенство Гамильтона — Кели. Как мы знаем, в общем случае

$$(A - \lambda_1 I)(A - \lambda_2 I) \dots (A - \lambda_n I) = 0. \quad (2-11.24)$$

Это тождество дает возможность выразить n -ю степень матрицы A через более низкие ее степени. Если все λ_i — различны, то не существует тождества более низкой степени, т. е. нельзя какую-нибудь степень A , меньшую чем n , выразить через низшие степени. Но в случае кратных корней положение иное — может случиться, что матрица A , имеющая кратное собственное число, удовлетворяет полиномиальному тождеству более низкой степени, чем тождество Гамильтона — Кели (и поэтому степень A , меньшая чем n , линейно выражается через низшие степени матрицы A). Пусть, например λ_1 — корень m -й кратности детерминантного уравнения (4.4). Можно показать, что уравнение Гамильтона — Кели примет в этом случае вид

$$(A - \lambda_1 I)^m (A - \lambda_2 I) \dots (A - \lambda_{n-(m-1)} I) = 0.$$

Может случиться, что при этом не существует тождества более низкой степени, которому удовлетворяет матрица A . Однако может представиться и такой случай, когда рассматриваемая матрица A удовлетворяет, помимо предыдущего тождества, также и «приведенному тождеству»

$$(A - \lambda_1 I)^{m-1} (A - \lambda_2 I) \dots (A - \lambda_{n-(m-1)} I) = 0,$$

или даже тождеству

$$(A - \lambda_1 I)^{m-2} (A - \lambda_2 I) \dots (A - \lambda_{n-(m-1)} I) = 0$$

и т. д. *).

В случае рассмотренных нами трех матриц собственное значение $\lambda = 1$ имеет кратность 3. Следовательно, мы заранее знаем, что полное равенство Гамильтона — Кели имеет вид

$$(A - I)^3 = 0, \quad (2-11.25)$$

и это равенство, несомненно, справедливо для все трех матриц. Однако может случиться, что уже

$$(A - I)^2 = 0, \quad (2-11.26)$$

или даже

$$(A - I)^1 = 0. \quad (2-11.27)$$

В нашем примере первая заданная матрица представляет собой не что иное, как единичную матрицу, и следовательно, действительно осуществляется случай (27). Здесь матричное тождество наинизшей степени имеет не третью, а только первую степень, а матрица A имеет три линейно независимых собственных вектора, соответствующих одному и тому же собственному числу 1. В этом проявляется следующее общее положение: если тождество наинизшей степени, которому удовлетворяет матрица A (имеющая собственное число λ_1 , кратности m),

*) Если A — симметрическая матрица, имеющая собственное число λ_1 m -й кратности, то она удовлетворяет тождеству

$$(A - \lambda_1 I)^m (A - \lambda_2 I) \dots (A - \lambda_{n-(m-1)} I) = 0.$$

содержит множитель $A - \lambda_1 I$ только в k -й степени ($k < m$), то существует не более чем $m - k + 1$ различных линейно независимых собственных векторов, имеющих λ_1 своим собственным числом *). Если остальные собственные числа $\lambda_2, \lambda_3, \dots, \lambda_{n-(m-1)}$ различны между собой (и отличны от λ_1), то матрица A имеет n линейно независимых собственных векторов («полноосная матрица»).

Для рассматриваемой первой матрицы $m = 3, k = 1, m - k + 1 = 3$, причем существует три линейно независимых собственных вектора (в других случаях, в соответствии с указанным общим положением, число их могло бы быть и меньшим).

Рассмотрим теперь вторую матрицу. Образует $A - I$ и затем $(A - I)^2$:

$$A - I = \begin{bmatrix} 0 & 1 & 2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad (A - I)^2 = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 2 & 0 & 0 \end{bmatrix} \circ \begin{bmatrix} 0 & 1 & 2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = 0.$$

Итак, $A - I$ не равно нулю, но

$$(A - I)^2 = 0.$$

Кратный корень $\lambda = 1$ входит в тождество низшей степени с кратностью 2, а количество линейно независимых векторов, соответствующих собственному числу $\lambda = 1$, как было установлено выше, равно только двум.

В соответствии со сформулированным ранее общим положением в случае, подобном рассматриваемому, существует не больше чем $m - 1$ различных линейно независимых собственных векторов (m — кратность корня λ_1), имеющих λ_1 своим собственным числом; может все же случиться, что число различных независимых собственных векторов, имеющих λ_1 собственным числом, будет равно только $m - 2$, либо даже $m - 3$ и т. д.

Теперь переходим к исследованию третьей матрицы. В этом случае

$$A - I = \begin{bmatrix} 0 & -1 & 3 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix},$$

$$(A - I)^2 = \begin{bmatrix} 0 & 0 & 0 \\ -1 & 0 & 0 \\ 3 & 2 & 0 \end{bmatrix} \circ \begin{bmatrix} 0 & -1 & 3 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & -2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

$$(A - I)^3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -2 & 0 & 0 \end{bmatrix} \circ \begin{bmatrix} 0 & -1 & 3 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = 0.$$

*) Более того, каждый вектор, являющийся произвольной линейной комбинацией этих собственных векторов, также является собственным вектором матрицы A , имеющим λ_1 своим собственным числом.

В этом случае кратный корень $\lambda = 1$ входит в матричное тождество низшей степени с кратностью 3, а число собственных векторов, соответствующих собственному числу $\lambda = 1$, равно только единице (как мы в этом убедились выше).

И в этом проявляется сформулированное ранее общее положение.

Таким образом, мы видим, что степень неполноосности матрицы зависит в некоторой мере от кратности, с которой сливающиеся корни входят в тождество низшей степени, связывающее степени матрицы *).

12. Инвариантность матричных равенств относительно ортогональных преобразований

Метод преобразования координат играет весьма важную роль при изучении матриц. Координаты не имеют абсолютного значения и могут быть заменены другими координатами. Мы можем рассматривать матрицу в рамках определенной системы отсчета, но можем также ввести и новую систему отсчета, которая окажется более соответствующей характеру рассматриваемой задачи. Вопрос о «наиболее адекватной системе координат» часто имеет поэтому большое значение. Мы можем начать с заданной системы координат, которая, возможно, не соответствует общим свойствам данной проблемы. Мы можем получить лучшую основу для нашего анализа, если перейдем от первоначальной системы к более адекватной системе. Однако в принципе *все* системы координат одинаково допустимы. Системы координат можно сравнить с лесами при возведении здания. Леса не принадлежат зданию, они служат только для того, чтобы иметь доступ ко всем участкам здания. И после того, как здание сооружено, леса могут быть сняты. Нельзя смешивать свойств, принадлежащих лесам, со свойствами самого здания. Сущность системы координат состоит в том, что она является *вспомогательной конструкцией*, дающей только возможность некоторого определенного *описания* природы, но не принадлежащей к ее подлинной *сущности*.

В матричной алгебре матрица ассоциируется с определенной пространственной фигурой, и анализ этой фигуры удобно выполнить, применяя координаты. Эти координаты не имеют абсолютного значения и могут быть заменены другими. Однако внутренние закономерности, выраженные в матричных уравнениях, не могут нарушаться случайным выбором системы координат, в которой эти матрицы рассматривались. Этот факт можно выразить и в следующей форме: уравнения матричной алгебры «инвариантны» относительно преобразования координат.

*) Полное исследование степени неполноосности несимметрической матрицы читатель найдет, например, в гл. VII книги [6]. Другое изложение дано в статье [12].

Но x и b являются оба векторами, и поэтому оба следуют одному и тому же закону преобразования (7):

$$x = U\bar{x}, \quad b = U\bar{b}. \quad (2-12.10)$$

Внеся эти выражения в (8), мы получаем

$$AU\bar{x} = U\bar{b}. \quad (2-12.11)$$

Умножим это равенство слева на $\tilde{U}$ и воспользуемся соотношениями ортогональности (6). Тогда

$$\tilde{U}AU\bar{x} = \bar{b}.$$

Следовательно, чтобы получить наше уравнение в виде (9), мы должны положить

$$\bar{A} = \tilde{U}AU. \quad (2-12.12)$$

Таким образом, преобразование матрицы отличается от преобразования вектора тем, что A умножается слева на $\tilde{U}$ и справа на U , тогда как при преобразовании вектора он умножается только слева на $\tilde{U}$. В самом деле, если равенство (7) умножим слева на $\tilde{U}$, то получим

$$\bar{x} = \tilde{U}x, \quad (2-12.13)$$

тогда как

$$\tilde{A} = \tilde{U}AU.$$

Рассмотрим теперь некоторые произвольные комбинации матриц и проверим, сохраняется ли закон преобразования (12):

$$\bar{A} + \bar{B} = \tilde{U}AU + \tilde{U}BU = \tilde{U}(A + B)U.$$

Итак, сумма матриц следует тому же закону преобразования, что и отдельная матрица. То же самое верно и относительно произведения матриц

$$\bar{A}\bar{B} = \tilde{U}AU\tilde{U}BU = \tilde{U}ABU,$$

в силу того, что $\tilde{U}U = I$.

Мы можем теперь сказать, что *любая* комбинация матриц, полученная с помощью основных операций сложения и умножения, преобразуется таким же точно образом, как и одна матрица. Это означает следующее. Пусть

$$F(A, B, \dots, P)$$

— какая-нибудь целая алгебраическая функция от произвольных матриц $A, B, \dots, P$. Тогда

$$F(\bar{A}, \bar{B}, \dots, \bar{P}) = \tilde{U}F(A, B, \dots, P)U.$$

Поэтому, если мы имеем произвольное алгебраическое матричное равенство

$$F(A, B, \dots, P) = 0, \quad (2-12.14)$$

то имеет место также и равенство

$$F(\bar{A}, \bar{B}, \dots, P) = 0. \quad (2-12.15)$$

Это показывает, что *матричные равенства остаются инвариантными при произвольных ортогональных преобразованиях.*

Мы можем сделать еще один шаг и включить в число допустимых операций также «транспонирование». Если

$$\bar{A} = \tilde{U}AU,$$

то, в силу правила транспонирования, мы имеем

$$\tilde{\bar{A}} = \tilde{U}\tilde{A}U.$$

Мы видим, что транспонированная матрица преобразуется точно так же, как и сама матрица. Следовательно, матричные равенства, образованные при помощи алгебраических операций и *транспонирования*, остаются инвариантными при ортогональных преобразованиях. Из равенства

$$F(A, B, \dots, P, \tilde{A}, \tilde{B}, \dots, \tilde{P}) = 0$$

следует

$$F(\bar{A}, \bar{B}, \dots, \bar{P}, \tilde{\bar{A}}, \tilde{\bar{B}}, \dots, \tilde{\bar{P}}) = 0.$$

Например, если $A - \tilde{A} = 0$, то также и $\bar{A} - \tilde{\bar{A}} = 0$, что означает, что симметрия матрицы сохраняется при ортогональных преобразованиях. Аналогично, если

$$A = -\tilde{A}, \text{ то и } \bar{A} = -\tilde{\bar{A}},$$

т. е., если матрица «антисимметрична» в одной ортогональной системе координат, то она остается антисимметрической во *всех* ортогональных системах координат.

Два основных понятия остаются неизменными при ортогональном преобразовании. Одно из них — единичная матрица

$$\bar{I} = \tilde{U}IU = \tilde{U}U = I,$$

другое — «скалярное произведение» (внутреннее произведение) двух векторов

$$\bar{x} \bar{y} = \tilde{U}x\tilde{U}y = y \cdot U\tilde{U}x = y \cdot x = x \cdot y.$$

Преобразование координат выражается следующими матричными равенствами:

$$\left. \begin{aligned} x &= U\bar{x}, \\ \xi &= V\bar{\xi}. \end{aligned} \right\} \quad (2-13.7)$$

Рассмотрим опять матричное равенство

$$Ax = b \quad (2-13.8)$$

и применим преобразование координат

$$x = U\bar{x}, \quad b = U\bar{b}.$$

Тогда

$$AU\bar{x} = U\bar{b},$$

и, умножив слева на $\tilde{V}$, получим

$$\tilde{V}AU\bar{x} = \bar{b}.$$

Это показывает, что преобразование матрицы нужно теперь производить согласно равенству

$$\bar{A} = \tilde{V}AU. \quad (2-13.9)$$

С другой стороны, если мы умножим слева первое равенство (7) на $\tilde{V}$, то получим

$$\bar{x} = \tilde{V}x.$$

Опять можно заметить, что преобразование матрицы отличается от преобразования вектора тем, что появляется дополнительный множитель справа.

Так как $\tilde{V}$ есть матрица, обратная матрице U , то мы можем также написать

$$\bar{A} = U^{-1}AU. \quad (2-13.10)$$

Кроме того, мы могли бы исходить в своих выводах из «сопряженного равенства»

$$\tilde{A}\xi = \beta, \quad (2-13.11)$$

которое является записью уравнения (8) в сопряженной системе координат. Тогда соотношения

$$\xi = V\bar{\xi}, \quad \beta = V\bar{\beta} \quad (2-13.12)$$

приводят к равенству

$$\tilde{\bar{A}} = \tilde{U}\tilde{A}V, \quad (2-13.13)$$

которое можно записать также и в форме

$$\tilde{A} = V^{-1} \tilde{A} V. \quad (2-13.14)$$

Равенство (13) находится в согласии с (9) и может быть получено из последнего с помощью транспонирования.

Мы снова можем, как и в § 12, распространить закон преобразования (9) на сумму и произведение матриц, а следовательно, и на любую комбинацию двух основных операций матричной алгебры — «сложения» и «умножения». При этом мы опять убедимся, что любое алгебраическое матричное равенство

$$F(A, B, \dots, P) = 0, \quad (2-13.15)$$

которое выполняется в одной системе координат, остается верным в любой другой системе

$$F(\bar{A}, \bar{B}, \dots, \bar{P}) = 0. \quad (2-13.16)$$

Мы снова доказали инвариантность матричных равенств относительно преобразования координат, распространив теперь этот факт на любые линейные преобразования.

Единственное отличие по сравнению со случаем ортогональных преобразований состоит в том, что равенство (15) не должно содержать операции транспонирования. Транспонированная матрица $\tilde{A}$ следует закону преобразования, отличному от закона преобразования матрицы A [ср. (10) и (14)]. Равенства матричных выражений, в которые входит транспонирование, *не* остаются инвариантными при произвольных преобразованиях координат¹⁾. Например, равенство

$$\tilde{A} = A$$

(которое выражает симметричность матрицы A) не сохраняется при произвольном линейном преобразовании, хотя оно остается верным при любых ортогональных преобразованиях.

Единичная матрица и скалярное произведение двух векторов также инвариантны относительно любого линейного преобразования

$$I = U^{-1} I U = U^{-1} U = I. \quad (2-13.17)$$

В случае скалярного произведения обязательно, чтобы сомножители задавались в *сопряженных представлениях*

$$\bar{x} \bar{\eta} = \tilde{V} x \tilde{U} \eta = \eta U \tilde{V} x = \eta x = x \eta. \quad (2-13.18)$$

¹⁾ Матрица A^{-1} , обратная матрице A , преобразуется так же, как и A . Поэтому равенство (15) может содержать обратные матрицы $A^{-1}, B^{-1}, \dots, P^{-1}$ (если они существуют).

14. Коммутативные и некоммутативные матрицы

Матричное умножение, вообще говоря, некоммутативно: $AB \neq BA$. Однако существуют матрицы, для которых выполняется коммутативный закон умножения:

$$AB = BA. \quad (2-14.1)$$

В силу принципа инвариантности матричных равенств относительно преобразования координат, равенство (1) должно оставаться справедливым в любой системе координат и, следовательно, должно выражать некоторое внутреннее свойство двух матриц A и B . Каково это свойство?

Мы видели, что, введя в качестве системы координат для полноосной матрицы ее главные оси, можно преобразовать ее в диагональную форму. Тогда в новой системе координат $\bar{A} = \Lambda$. Может случиться, что другая полноосная матрица B имеет те же главные оси, хотя и другие собственные значения, и следовательно, также окажется диагональной в этой новой системе координат. Пусть тогда

$$\bar{A} = \Lambda, \quad \bar{B} = \Lambda'.$$

Но произведение двух диагональных матриц есть опять диагональная матрица, а именно:

$$\Lambda\Lambda' = \begin{bmatrix} \lambda_1\lambda'_1 & 0 & \dots & 0 \\ 0 & \lambda_2\lambda'_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_n\lambda'_n \end{bmatrix}.$$

Элементы этой матрицы симметричны относительно обоих сомножителей и, следовательно,

$$\Lambda\Lambda' = \Lambda'\Lambda.$$

Итак, *диагональные матрицы всегда коммутативны*. Но тогда A и B коммутативны в этой новой системе координат, и равенство (1) доказано для этой частной системы. Так как, однако, была уже доказана независимость такого равенства от системы координат, то мы тем самым нашли достаточное условие коммутативности двух матриц, а именно, *что их главные оси параллельны*. Можно доказать, что оно является (для полноосных матриц) также и необходимым.

Например, следующие две полноосные матрицы коммутативны, и поэтому имеют одни и те же главные оси:

$$A = \begin{bmatrix} 33 & 16 & 72 \\ -24 & -10 & -57 \\ -8 & 4 & -17 \end{bmatrix}, \quad B = \begin{bmatrix} -47 & -32 & -84 \\ 36 & 24 & 66 \\ 12 & 8 & 22 \end{bmatrix}.$$

Единичная матрица геометрически представляется сферой. Любая ось может быть выбрана в качестве главной оси. Поэтому единичная матрица перестановочна с любой матрицей

$$IA = AI = A.$$

Далее матрица A^{-1} , обратная матрице A , имеет те же главные оси, что и A . Следовательно A и A^{-1} коммутативны. Матрицы, имеющие различные главные оси, не могут быть коммутативны. Матричное умножение, вообще говоря, некоммутативно по той причине, что две матрицы в общем случае имеют произвольно ориентированные главные оси. Если же главные оси матриц сомножителей одинаковы, то умножение становится коммутативным.

15. Обращение матриц. Гауссов метод исключения

Система линейных алгебраических уравнений может быть записана в следующей матричной форме:

$$Ax = b, \quad (2-15.1)$$

где A — заданная матрица, b — данный вектор, а x — искомый вектор. Это уравнение можно формально решить, умножив его слева на A^{-1} :

$$x = A^{-1}b. \quad (2-15.2)$$

Матрица A^{-1} называется «обратной» к матрице A , а операция вычисления A^{-1} называется обращением матрицы A .

Не всегда, однако, необходимо действительно обратить матрицу для решения линейного уравнения (1). Это зависит от той роли, которую играет правая часть уравнения (1) в рассматриваемой проблеме. Возможно, что правая часть есть точно такая же существенная часть нашей задачи, как и матрица A . В этом случае нас вполне удовлетворяет возможность решения уравнения (1) при данном частном значении правой части. Однако часто бывает, что правая часть уравнения (1) имеет более *случайный* характер. Матрица A существенно связана с данным физическим процессом, тогда как правая часть часто имеет значение «силовой функции» и нам может понадобиться узнать поведение данной конструкции при различных силовых функциях. Поэтому придется свободно менять правую часть, хотя левая часть уравнения будет оставаться неизменной. В этом случае нам действительно нужно знать A^{-1} , так как тогда, подействовав матрицей A^{-1} на b , мы сразу получим искомое решение. Если мы не знаем A^{-1} , то для каждого отдельного значения b приходится проделывать большое число алгебраических операций. С другой стороны, построение матрицы, обратной к A , требует гораздо большего труда, чем решение системы линейных алгебраических уравнений для заданного значения правой части.

Основной метод обращения матрицы был дан Гауссом и называется гауссовым методом исключения. Он пригоден как для решения системы линейных уравнений, так и для обращения матрицы. Принцип этого метода прост, и связанные с ним операции легко выполнимы. Он принадлежит к тому классу вычислительных приемов, в которых сравнительно небольшое число довольно сложных операций заменяется большим числом весьма простых операций. Этот метод дает возможность постепенно перемещать нулевые элементы единичной матрицы на левую сторону, тогда как с правой стороны матрица все более и более заполняется новыми элементами*). К концу процесса единичная матрица оказывается слева, а правая сторона заполнена искомыми элементами. Этим заканчивается работа.

Мы рассмотрим этот процесс, применив его сначала к решению линейной системы уравнений с заданной правой частью. Запишем уравнение (1) в форме

$$Ax - b = 0 \quad (2-15.3)$$

и представим соответствующую систему уравнений символически, выписав только ее коэффициенты; соответствующие переменные поместим в верхней строке схемы

$$\begin{array}{cccccc} x_1 & x_2 & \dots & x_n & - & 1 \\ \hline a_{11} & a_{12} & \dots & a_{1n} & b_1 & \\ a_{21} & a_{22} & \dots & a_{2n} & b_2 & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \dots & a_{nn} & b_n & \end{array} = 0. \quad (2-15.4)$$

Так как каждое уравнение можно умножить на какое-либо число, то мы можем умножить на произвольное число любую строку схемы. Кроме того, прибавление к уравнению любого другого уравнения этой системы дает систему, эквивалентную первоначальной. Соответственно этому мы можем умножить любую строку указанной выше схемы на какое-либо число и прибавить полученную строку к любой другой строке этой схемы. Это и есть две основные операции, на которых основан метод исключения.

Сначала находим наибольший по абсолютной величине элемент матрицы. Пусть он будет a_{ik} . Затем делим все элементы i -й строки на a_{ik} , что дает новое значение для элемента a_{ik} , равное единице. Наша цель теперь сделать все остальные элементы k -го столбца равными нулю (тогда k -й столбец будет состоять из нулей и только одной единицы). Для того чтобы какой-либо определенный элемент этого столбца сделать равным нулю, мы умножаем на него i -ю строку

*) Это несколько фигуральное выражение имеет в виду, что уравнение (1) можно записать в форме

$$Ax = Ib,$$

которая последовательными этапами превращается в форму

$$Ix = A^{-1}b.$$

и вычитаем полученную строку из той строки, в которой этот элемент расположен. Рассмотрим пример:

$$\begin{bmatrix} x_1 & x_2 & x_3 & -1 \\ 33 & 16 & 72 & 359 \\ -24 & -10 & -57 & -281 \\ -8 & -4 & -17 & -85 \end{bmatrix} = 0.$$

Наибольший элемент 72 матрицы находится в первой строке. Разделив эту строку на 72, получаем новую первую строку

$$0,458333 \quad 0,222222 \quad 1 \quad 4,98611$$

Единица появилась в третьем столбце. Мы превратим элементы этого столбца, -57 и -17 , в нули, умножая новую первую строку на 57 и добавляя полученную строку ко второй строке и аналогично умножая на 17 и добавляя результат к третьей строке. После этого матрица примет следующий вид:

$$\begin{bmatrix} x_1 & x_2 & x_3 & -1 \\ 0,458333 & 0,222222 & 1 & 4,98611 \\ 2,12498 & 2,66665 & 0 & 3,20827 \\ -0,208339 & -0,222226 & 0 & -0,23613 \end{bmatrix}. \quad (\alpha)$$

Переменная x_3 теперь «исключена» — второе и третье уравнения не содержат ее более. Первоначальная система из трех уравнений с тремя неизвестными содержит в себе систему из двух уравнений с двумя неизвестными. В общем случае сведение к нулю элементов одного столбца приводит первоначальную систему n уравнений с n неизвестными к системе, состоящей из $(n-1)$ уравнений с $(n-1)$ неизвестными.

Продолжим решение нашего примера. Отыщем наибольший элемент в оставшейся матрице, содержащей две строки и два столбца. Он равен 2,66665. Делим на него вторую строку матрицы (α) и получаем новую вторую строку

$$0,796872 \quad 1 \quad 0 \quad 1,203108.$$

Теперь приведем к нулю второй столбец, умножив эту строку последовательно на 0,222222 и $-0,222226$ и вычитая ее соответственно из первой и третьей строки. Мы получим новую матрицу

$$\begin{array}{cccc} 0,281250 & 0 & 1 & 4,71875 \\ 0,796872 & 1 & 0 & 1,20311 \\ -0,031253 & 0 & 0 & 0,031232. \end{array}$$

Разделим наконец третью строку на $-0,031253$; и мы получим

$$\begin{array}{cccc} 1 & 0 & 0 & -1,00067. \end{array}$$

Приводим к нулю первый столбец, умножая новую третью строку на $0,281250$ и $0,796872$ и вычитая результат соответственно из первой и второй строки. Это приводит нас к матрице

$$\begin{bmatrix} 0 & 0 & 1 & 5,00019 \\ 0 & 1 & 0 & 2,00052 \\ 1 & 0 & 0 & -1,00067 \end{bmatrix}.$$

Преобразованная система уравнений имеет, таким образом, вид

$$\begin{array}{rcl} & x_3 & - 5,00019 = 0, \\ & x_2 & - 2,00052 = 0, \\ x_1 & & + 1,00067 = 0. \end{array}$$

Это дает решение

$$\begin{array}{rcl} x_1 & = & -1,00067 \quad [\text{точно: } -1], \\ x_2 & = & 2,00052 \quad [\text{точно: } 2], \\ x_3 & = & 5,00019 \quad [\text{точно: } 5]. \end{array}$$

Совершенно таким же образом проводится операция обращения матрицы. Единственное различие лишь в том, что с правой стороны теперь должна стоять матрица I . Следовательно, нам приходится оперировать с n столбцами вместо одного. Однако последовательные операции при этом остаются теми же, что и в предыдущем алгоритме: вместо матрицы (α) имеем матрицу

$$\begin{bmatrix} 33 & 16 & 72 & 1 & 0 & 0 \\ -24 & -10 & -57 & 0 & 1 & 0 \\ -8 & -4 & -17 & 0 & 0 & 1 \end{bmatrix}.$$

Первое преобразование приведет к матрице

$$\begin{bmatrix} 0,45833 & 0,222222 & 1 & 0,0138888 & 0 & 0 \\ 2,12498 & 2,66665 & 0 & 0,791662 & 1 & 0 \\ -0,208339 & -0,222226 & 0 & 0,236110 & 0 & 1 \end{bmatrix}.$$

Второе преобразование приведет к матрице

$$\begin{bmatrix} 0,281250 & 0 & 1 & -0,052083 & -0,083333 & 0 \\ 0,796872 & 1 & 0 & 0,296875 & 0,375002 & 0 \\ -0,031253 & 0 & 0 & 0,302083 & 0,083395 & 1 \end{bmatrix}.$$

Третье преобразование даст

$$\begin{bmatrix} 0 & 0 & 1 & 2,66640 & 0,66661 & 8,9991 \\ 0 & 1 & 0 & 7,9992 & 2,49983 & 25,4974 \\ 1 & 0 & 0 & -9,66572 & -2,66646 & -31,9969 \end{bmatrix}.$$

Положение единиц в первой части последней матрицы показывает, что строки в окончательном результате должны быть переставлены в порядке 3, 2, 1. Таким образом, обратная матрица принимает вид

$$A^{-1} = \begin{bmatrix} -9,66572 & -2,66646 & -31,9969 \\ 7,9992 & 2,49983 & 25,4974 \\ 2,66640 & 0,66661 & 8,9991 \end{bmatrix}.$$

Точное значение обратной матрицы, полученное другим способом с пятью десятичными знаками, есть

$$A^{-1} = \begin{bmatrix} -9,66667 & -2,66667 & -32,00000 \\ 8,00000 & 2,50000 & 25,50000 \\ 2,66667 & 0,66667 & 9,00000 \end{bmatrix}.$$

Расхождения между этими двумя значениями матрицы A вызваны ошибками округления. Уже этот простой пример матрицы третьего порядка показывает, как быстро накапливаются ошибки округления. Хотя вычисления производились с точностью до 10^{-6} , результат получился с точностью лишь до 10^{-4} . При матрицах более высокого порядка это накопление ошибок округления может оказаться весьма серьезным.

16. Последовательная ортогонализация матрицы

Вычислительная сторона задачи обращения матрицы совершенно отлична от чисто математической. Наши вычисления всегда имеют ограниченную точность. Ошибки округления постоянно накапливаются и ввиду большого числа операций, которые приходится производить, они могут в конце концов исказить искомые результаты. Гауссов метод исключения обеспечивает надлежащие результаты, если можно пренебречь ошибками округления. Но при обращении матриц высокого порядка этого почти никогда не бывает. А между тем именно эти задачи представляют большой интерес в связи с тем, что много проблем современной физики и техники приводят к решению систем линейных уравнений с очень большим числом неизвестных. Какие методы могут быть применены в этих условиях?

Изобретение электронных вычислительных машин открыло новую главу в истории численного анализа. Исключительная быстрота, с которой эти машины выполняют основные арифметические операции,

приводит к совершенно новому взгляду на численные расчеты. Главное внимание теперь уже направлено не на то, чтобы получить результат при наименьшем числе производимых операций. Более важным представляется вопрос простоты кодирования алгоритма задачи вместе с требованием большой точности. В задаче обращения матрицы нас интересует такой алгоритм, который, несмотря на ограниченную точность арифметических операций, не может привести к неудаче, независимо от того, насколько велик порядок матрицы.

При рассмотрении вопроса о точности следует иметь в виду, что ни при каких обстоятельствах мы не можем ожидать *абсолютной* точности. Точно так же не может быть речи о получении результатов с произвольно установленной точностью. Элементы данной матрицы A редко бывают математически точными числами. В большинстве случаев элементы a_{ik} матрицы получены при помощи некоторых *измерений*, которые уже в силу одного этого обстоятельства имеют ограниченную точность. Допустим, что в результате некоторого процесса обращения мы получим приближенно матрицу, обратную A . Эта матрица $\bar{A}^{-1}$ является точным обращением некоторой матрицы $\bar{A}$, которая отличается от A на малое слагаемое αA_1 :

$$\bar{A} = A + \alpha A_1.$$

Допустим также, что все элементы матрицы αA_1 меньше, чем возможные ошибки определения элементов a_{ik} . При этих условиях мы можем считать $\bar{A}$ «численно эквивалентной» матрице A и полученную $\bar{A}^{-1}$ принять за правильный ответ нашей задачи. Было бы совершенно напрасно корректировать $\bar{A}^{-1}$ с целью получить точное значение A^{-1} . Математически точный ответ не имеет никакого значения ввиду ограниченной точности заданной матрицы A . В случае задания «математически точной» матрицы A , конечно, теряет смысл замена ее «численно эквивалентной» матрицей $\bar{A}$. Однако даже в этом случае целесообразно заменить элементы матрицы A числами, взятыми с точностью до определенного десятичного знака, провести весь процесс обращения и в конце, в случае надобности, скорректировать $\bar{A}^{-1}$ методом возмущений (ср. § 23).

Метод обращения, описываемый в настоящем параграфе, имеет то достоинство, что на протяжении всего процесса мы можем контролировать ошибки округления. Точность вычислений определяется точностью элементов матрицы A и может быть, в случае надобности, отрегулирована даже до переменной точности. Мы разберем сначала математическую сторону этого метода, а затем обратимся к вопросу об ошибках округления.

Рассмотрим снова уравнение

$$Ax = b, \quad (2-16.1)$$

В одном частном случае задача эта легко разрешается. А именно может случиться, что базисные векторы взаимно *ортогональны* и длины их равняются 1:

$$u_i u_k = 0 \quad (i \neq k), \quad u_i^2 = 1. \quad (2-16.6)$$

Тогда эта система самосопряженна, и мы получаем

$$v_i = u_i. \quad (2-16.7)$$

Обратная матрица поэтому есть просто матрица, транспонированная с первоначальной

$$A^{-1} = \tilde{A}. \quad (2-16.8)$$

Таким образом, ортогональные системы координат обладают особым преимуществом при операциях с системами линейных уравнений. Мы постараемся поэтому ввести вспомогательную систему координат, обладающую свойством ортогональности, решим задачу в этой новой системе и затем вернемся к первоначальной системе.

Выбираем первый вектор u_1 в качестве первого вектора новой системы. При этом, однако, мы нормируем u_1 , приведя его длину к 1:

$$w_1 = \frac{u_1}{\sqrt{u_1^2}}. \quad (2-16.9)$$

Следующий вектор w_2 мы берем в плоскости, определяемой векторами u_1 и u_2 *). В этой плоскости мы построим вектор, ортогональный вектору w_1 и имеющий длину 1. Этот вектор мы и назовем w_2 ; он определяется с точностью до знака. Затем мы переходим в пространство, порождаемое *тремя* исходными векторами u_1, u_2, u_3 **). В этом пространстве мы строим вектор w_3 , который ортогонален как вектору w_1 , так и вектору w_2 и имеет длину 1; он также определится с точностью до знака. Далее мы переходим в пространство, порождаемое *четырьмя* первыми векторами u_1, u_2, u_3, u_4 , и находим вектор w_4 длины 1 и ортогональный трем векторам w_1, w_2, w_3 . Мы будем последовательно выполнять такие построения, пока, наконец, не будет исчерпано пространство всех векторов $u_1, u_2, \dots, u_n$. Этот процесс приводит нас к системе n взаимно ортогональных векторов $w_1, w_2, \dots, w_n$, имеющих длину 1.

Каждый следующий вектор получается в два последовательных этапа. Сначала мы строим вектор

$$w'_i = u_i - p_{i1}w_1 - p_{i2}w_2 - \dots - p_{i,i-1}w_{i-1}, \quad (2-16.10)$$

где коэффициенты p_{ik} определяются из условия, что w'_i перпендику-

*) Следовательно, и в плоскости векторов w_1 и u_2 .

**) Следовательно, и тремя векторами w_1, w_2, u_3 .

лярен всем ранее построенным векторам w_k (в предположении, что эти последние взаимно ортогональны и нормированы). Поэтому скалярное произведение w'_i на какой-либо вектор w_k дает

$$p_{ik} = u_i w_k. \quad (2-16.11)$$

Затем мы нормируем вектор ¹⁾

$$w_i = \frac{w'_i}{|w'_i|}. \quad (2-16.12)$$

Обозначив

$$p_{ii} = |w'_i|, \quad (2-16.13)$$

получаем, согласно формуле (10),

$$u_i = p_{i1}w_1 + p_{i2}w_2 + \dots + p_{ii}w_i. \quad (2-16.14)$$

Коэффициенты p_{ik} ($k \leq i$) можно рассматривать как элементы «треугольной матрицы», т. е. матрицы, которая имеет отличные от нуля элементы по диагонали и ниже диагонали, тогда как все элементы, расположенные выше диагонали, равны нулю.

Равенства (14) могут быть теперь записаны в матричной форме

$$A = W\tilde{P}, \quad (2-16.15)$$

где W есть матрица, столбцы которой представляют последовательные векторы $w_1, w_2, \dots, w_n$. По построению W есть ортогональная матрица

$$\tilde{W}W = I \quad (2-16.16)$$

и, следовательно,

$$W^{-1} = \tilde{W}. \quad (2-16.17)$$

Ниже будет показано, что треугольная матрица P может быть обращена путем последовательного исключения (ср. § 17), причем получится опять треугольная матрица

$$Q = P^{-1}. \quad (2-16.18)$$

Но тогда мы, в силу равенства (15), получаем

$$A^{-1} = \tilde{Q}\tilde{W}. \quad (2-16.19)$$

Тем самым задача обращения матрицы A решена.

¹⁾ Случай деления на нуль не может встретиться, так как вектор w'_i не может иметь сплошь нулевые координаты, если определитель матрицы A не равен нулю. Действительно, из равенства $w'_i = 0$ мы получили бы, заменив все w_k их выражениями через векторы $u_1, u_2, \dots, u_k$, равенство вида

$$0 = u_i + \lambda_1 u_1 + \lambda_2 u_2 + \dots + \lambda_{i-1} u_{i-1},$$

что противоречит условию линейной независимости базисных векторов.

а затем исправляем их по следующей схеме:

$$p_{ik} = p'_{ik} - (\varepsilon_{k1}p'_{i1} + \dots + \varepsilon_{k,i-1}p'_{i,i-1}). \quad (2-16.23)$$

Получив окончательные значения p_{ik} , мы вновь образуем вектор w'_i по формуле (10) и нормируем его длину согласно формуле (12). Значение p_{ii} мы опять определяем с помощью равенства (13) без дальнейших исправлений.

Ортогональность системы векторов w_i в пределах выбранного числа десятичных знаков теперь обеспечивается самим методом построения каждого нового вектора. Важно, однако, чтобы при нахождении величины p_{ik} мы удерживали достаточное число десятичных знаков. При весьма косоугольных системах длина вектора w'_i может оказаться очень малой, ибо мы можем потерять несколько десятичных знаков при построении w'_i . Мы должны восполнить эту потерю добавлением соответствующего числа значащих цифр при записи p_{ik} , не требуя, однако, увеличенной точности ранее полученных векторов w_k . Поэтому если даже в вычислениях возникает случайная ошибка, то нет необходимости повторять все вычисления с повышенной точностью. Самое худшее, что может случиться (если исключить системы, столь близкие к особым, что для них требуется *точное* решение уравнений (20); в этом случае система лишена какого-либо физического значения) это то, что для величин (22) и (23) может потребоваться максимум точности. В предвидении этой возможности целесообразно получить набор значений ε_{ik} со *всею точностью*.

Число десятичных знаков, которыми можно ограничиться при вычислении векторов w_i , устанавливается следующим образом. Мы интерпретируем отброшенную часть вектора w'_i , определенного по формулам (10) и (12) как изменение вектора u_i , который составляет i -й столбец заданной матрицы. Если вектор w_i взят с точностью до μ десятичных знаков, то ошибки округления в любой координате

не могут превысить $\frac{1}{2} 10^{-\mu}$. Длина вектора w'_i не может превышать

длины u_i . Следовательно, вектор $\tilde{u}_i$, который занимает теперь место вектора u_i (в результате округления) в нашем численном процессе обращения не может отличаться от u_i на величину, большую чем $0,5 \cdot 10^{-\mu}$ -я часть длины u_i . Отсюда следует, что если точность элементов i -го столбца матрицы A после деления их на длину этого столбца может быть гарантирована до μ десятичных знаков, то можно при вычислении w_i ограничиться μ десятичными знаками. Мы, таким образом, пришли к замене заданной матрицы A другой, но численно ей эквивалентной матрицей $\bar{A}$, которая разложена в произведение двух множителей по формуле (15). Теперь следует учесть, что матрица W не вполне ортогональна, и поэтому ее обращение не вполне совпадает с $\tilde{W}$. Так как, однако, мы уже вычислили матрицу ε

треугольнике), который расположен в одном с ним столбце, прекращая это поэлементное умножение, когда в верхней строке достигаем диагонального элемента. Следовательно, для образования элемента q_{53} мы будем иметь лишь два члена:

$$q_{55}p_{53} + q_{54}p_{43},$$

так как следующий член содержал бы элемент q_{53} , который еще отсутствует в схеме. Результирующую сумму составленных произведений умножаем на q_{33} (верхний диагональный элемент) и в полученном результате меняем знак

$$q_{53} = -(q_{55}p_{53} + q_{54}p_{43})q_{33}.$$

Поясним этот алгоритм на простом примере. Пусть нужно обработать следующую треугольную матрицу:

$$P = \begin{bmatrix} 2 & & & \\ 3 & 5 & & \\ 0 & -7 & 10 & \\ 6 & 1 & 8 & 1 \end{bmatrix}.$$

Строим вспомогательную схему

$$\begin{array}{cccc} 0,5 & 3 & 0 & 6 \\ -0,3 & 0,2 & -7 & 1 \\ -0,21 & 0,14 & 0,1 & 8 \\ -1,02 & -1,32 & -0,8 & 1. \end{array}$$

Элементы этой схемы получены следующим образом:

элемент (2,1) — умножаем 0,2 на 3, затем на 0,5 и в результате меняем знак на минус:

$$-0,2 \cdot 3 \cdot 0,5 = -0,3,$$

элемент (4,2) — умножаем 1 на 1; $-0,8$ на -7 , сумму этих произведений умножаем на 0,2 и меняем в произведении знак:

$$-[1 \cdot 1 + (-0,8) \cdot (-7)] \cdot 0,2 = -1,32$$

и так далее.

Таким образом, получаем искомую матрицу Q :

$$Q = \begin{bmatrix} 0,5 & & & \\ -0,3 & 0,2 & & \\ -0,21 & 0,14 & 0,1 & \\ -1,02 & -1,32 & -0,8 & 1 \end{bmatrix}.$$

Наши вычисления мы проверим, образовав произведение PQ , которое должно равняться единичной матрице I . Для того чтобы умножение можно было производить столбец на столбец, транспонируем первый множитель

$$\begin{bmatrix} 2 & 3 & 0 & 6 \\ 0 & 5 & -7 & 1 \\ 0 & 0 & 10 & 8 \\ 0 & 0 & 0 & 1 \end{bmatrix} \circ \begin{bmatrix} 0,5 & 0 & 0 & 0 \\ -0,3 & 0,2 & 0 & 0 \\ -0,21 & 0,14 & 0,1 & 0 \\ -1,02 & -1,32 & -0,8 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

На этом построение матрицы, обратной заданной треугольной матрице, заканчивается.

18. Численный пример последовательной ортогонализации матрицы

Приведенная ниже матрица A пятого порядка была получена в результате решения одной производственной задачи методом наименьших квадратов. Матрицы, получаемые таким путем, всегда симметричны и имеют еще следующее свойство: они «положительно определены». Это значит, что квадратичная форма, соответствующая матрице: $\sum a_{ik}x_i x_k$, не может равняться нулю или отрицательному числу ни при каких значениях переменных (за исключением тривиального случая, когда все x_i равны нулю).

Такие матрицы можно надлежащим образом нормировать следующим приемом. Преобразуем переменные x_i , меняя масштаб

$$x'_i = \sqrt{a_{ii}} x_i, \quad (2-18.1)$$

и делим i -е уравнение на $\sqrt{a_{ii}}$. Тогда элементы матрицы преобразуются следующим образом:

$$a''_{ik} = \frac{a_{ik}}{\sqrt{a_{ii}} \sqrt{a_{kk}}}. \quad (2-18.2)$$

В результате этого преобразования диагональные элементы матрицы оказываются равными 1. Кроме того, все недиагональные элементы лежат между ± 1 *). В нашей задаче такое начальное нормирование (2) было выполнено заранее, и мы начинаем прямо с нормированной матрицы. Элементы этой матрицы заданы с точностью до пятого десятичного знака, соответственно точности наблюдений, из

*) Этот факт является следствием положительной определенности матрицы.

которых были получены эти элементы:

$$A = \begin{bmatrix} 1 & 0,99877 & 0,99297 & 0,98341 & 0,99580 \\ 0,99877 & 1 & 0,99325 & 0,98044 & 0,99363 \\ 0,99297 & 0,99325 & 1 & 0,98233 & 0,98776 \\ 0,98341 & 0,98044 & 0,98233 & 1 & 0,98749 \\ 0,99580 & 0,99363 & 0,98776 & 0,98749 & 1 \end{bmatrix}.$$

Тот факт, что все элементы близки к единице, указывает, что система весьма косоугольна.

Для обеспечения того, чтобы «эквивалентная матрица» не отклонялась от заданной матрицы больше чем на $\frac{1}{2}$ единицы пятого десятичного знака, при образовании матрицы W удержим шесть десятичных знаков. Определение элементов p_{ik} производилось с точностью до девяти десятичных знаков, так как вычисление векторов w_i' показало, что два десятичных знака точности терялись в процессе образования разностей (16.10). Поэтому постоянно требовались поправки (16.23). Нижеследующая таблица содержит в столбцах векторы w_i в том порядке, в котором они появлялись последовательно в процессе ортогонализации и нормирования:

$$10^6 W = \begin{bmatrix} 449819 & -074358 & -321824 & -476554 & -679308 \\ 449266 & 636326 & -372075 & -502115 & 051862 \\ 446657 & 360073 & 744406 & -317396 & 126348 \\ 442357 & -582393 & 297446 & 582606 & -192965 \\ 447930 & -347456 & -339669 & -283912 & 694732 \end{bmatrix}.$$

(Эта таблица была получена постепенно, но здесь выписана полностью; для удобства записи все векторы w_i умножены на 10^6 .) Почленное перемножение столбцов матрицы W между собой показало, что отклонение от ортогональности никогда не достигало 10^{-6} . Следующая таблица дает матрицу соответствующих произведений (ϵ -матрица текста), умноженных на 10^{12} :

$$10^{12} \epsilon = \begin{bmatrix} 547515 & -464906 & 338988 & 148894 & 292931 \\ -464906 & -265846 & -742034 & -026372 & 290133 \\ 338988 & -742034 & -062086 & 518499 & -317668 \\ 148874 & -026372 & 518499 & -816063 & -035420 \\ 292931 & 290133 & -317668 & -035420 & 886061 \end{bmatrix}.$$

Элементы P -матрицы появлялись строка за строкой. Кроме диагональных элементов — которые были получены путем извлечения квадратного корня из суммы квадратов компонент векторов w_i — каждый другой элемент p_{ik} дан в приведенной ниже таблице в виде суммы двух слагаемых, а именно предварительного значения p'_{ik} , полученного по формуле (16.22), и поправки, содержащейся во втором члене формулы (16.23), которая в таблице приводится непосредственно под

соответствующим значением p'_{ik} :

$$p_{11} = 2,223116$$

$$p_{21} = 2,220954971,$$

$$\text{попр.: } -1216,$$

$$p_{2k} = 2,220953755, \quad 0,003458902,$$

$$2,216535115, \quad 0,002963282,$$

$$\text{попр.: } -1212, \quad +1031,$$

$$p_{3k} = 2,216533903, \quad 0,002964313, \quad 0,01195886,$$

$$2,206282826, \quad 0,021036753, \quad 0,011995689,$$

$$\text{попр.: } -1222, \quad +1029, \quad -763,$$

$$p_{4k} = 2,206281604, \quad -0,021035724, \quad 0,011995689, \quad 0,01410346,$$

$$2,220276968, \quad -0,008670650, \quad -0,000826800, \quad 0,002258580,$$

$$\text{попр.: } -1220, \quad +1029, \quad -760, \quad -328,$$

$$p_{5k} = 2,220275748, \quad -0,008669621, \quad -0,000827560, \quad 0,002258252, \quad 0,004058572$$

Из приведенных вычислений можно сделать интересное заключение относительно *определителя* заданной матрицы A . Равенство (16.15) показывает, что определитель матрицы A равен произведению определителей W и P . Но определитель ортогональной матрицы W может равняться только ± 1 , причем знак определяется знаками диагональных элементов. В нашем случае диагональные элементы матрицы W положительны, следовательно, знак определителя — плюс. Определитель P — ввиду треугольного вида матрицы — есть просто произведение всех ее диагональных элементов. Это произведение в нашей задаче составляет $\Delta = 5,26367 \cdot 10^{-9}$, что показывает исключительно косоугольный характер данной линейной системы.

Теперь мы переходим к обращению матрицы P в соответствии с численной схемой § 17. Это дает матрицу Q :

$$\begin{bmatrix} 0,449819082 & & & & & \\ -288,827893 & 289,109087 & & & & \\ -11,7789627 & -71,6631707 & 83,6200106 & & & \\ -491,144346 & 492,1677422 & -71,1229472 & 70,9045865 & & \\ -592,171546 & 329,1112848 & 56,6243775 & -39,4524044 & 246,392080 & \end{bmatrix}$$

(все невыписанные элементы равны нулю). Наконец, формула (16.19) показывает, что произведение матрицы $\tilde{Q}$ на матрицу $\tilde{W}$ дает искомую обращенную матрицу A^{-1} :

$$A^{-1} = \begin{bmatrix} 661,793 & -456,547 & -31,482 & -6,990 & -167,376 \\ -456,526 & 474,825 & -63,888 & 33,556 & 12,778 \\ -31,499 & -63,878 & 91,976 & -27,490 & 31,131 \\ -6,970 & 33,542 & -27,491 & 48,922 & -47,545 \\ -167,401 & 12,800 & 31,128 & -47,539 & 171,176 \end{bmatrix}.$$

Точное обращение матрицы A должно было бы привести к симметричной матрице. Найденная обратная матрица несколько уклоняется от симметрии. Нет надобности, однако, стараться устранить это отклонение, так как мы знаем, что наша матрица A^{-1} является точным обращением некоторой матрицы, элементы которой отличаются от элементов заданной матрицы меньше чем на $\frac{1}{2}$ пятого десятичного знака (такие величины отклонений остаются вне досягаемости наших измерений). Любые заключения, содержащие величины столь малого порядка, были бы тем самым уже не реальны. Тот факт, что неточности появляются уже во втором десятичном знаке, является не ошибкой процесса обращения, а следствием очень косоугольного характера матрицы A . Столбцы матрицы A были даны с относительной точностью примерно $1 \cdot 10^{-6}$. В столбцах матрицы A^{-1} мы не можем ожидать относительной точности, превышающей $5 \cdot 10^{-4}$. На самом же деле отклонения от симметрии значительно меньше допустимого максимума ошибок.

19. Триангуляризация матрицы

Процесс ортогонализации, рассмотренный в § 16, имеет еще другой аспект. Если мы протранспонируем равенство (16.15) и умножим слева старое равенство на новое, то получим

$$\tilde{A}A = P\tilde{P}. \quad (2-19.1)$$

В этом равенстве ортогональная матрица W *полностью выпала*. Отсюда следует, что мы можем получить треугольную матрицу P , даже не строя явно ортогональных векторов w_i . Хотя этот прием, называемый триангуляризацией, намного короче, чем предыдущий, он имеет тот недостаток, что мы не можем столь просто держать под контролем ошибки округления, как в первом. Однако метод триангуляризации имеет свое преимущество, если матрица A не слишком косоугольна и если мы можем оперировать с достаточным числом дополнительных десятичных знаков, чтобы держать ошибки округления ниже опасного уровня.

Если записать уравнение (1) в виде системы линейных алгебраических уравнений, то мы увидим, что оно может быть решено путем перехода от строки к строке, а в каждой строке — слева направо. Каждое следующее уравнение привносит одно новое неизвестное, которое, таким образом, может быть исключено. К концу процесса будет определена полностью матрица P . Затем мы снова обращаем P и получаем новую треугольную матрицу Q . Но равенство (16.15) показывает, что можно исключить матрицу W , если умножить это равенство справа на $\tilde{Q}$:

$$W = A\tilde{Q} \quad (2-19.2)$$

и подставить затем это значение W в (16.19). Тогда получим

$$A^{-1} = \tilde{Q}Q\tilde{A}. \quad (2-19.3)$$

Матрица $\tilde{A}A$ по своему строению является симметрической и положительно определенной матрицей. Мы можем рассматривать равенство (1) как выражение того факта, что всякая симметрическая и положительно определенная матрица может быть представлена как произведение некоторой треугольной матрицы на транспонированную к ней матрицу. Но во многих проблемах прикладной математики исходная матрица A как раз и является симметрической и положительно определенной. В таком случае нет надобности умножать ее на $\tilde{A}$ и можно сразу же положить

$$A = P\tilde{P}. \quad (2-19.4)$$

Мы при этом не только освобождаемся от одного перемножения матриц — весьма трудоемкой операции в случае матриц высокого порядка, — но получаем еще то большое преимущество, что матрица P будет гораздо менее косоугольной, чем была первоначальная матрица. Произведение диагональных элементов матрицы P тогда представляет собой лишь *квадратный корень* первоначального определителя. Следовательно, потеря в значащих цифрах теперь намного менее ярко выражена, чем прежде. Это приведение симметрической матрицы к треугольной весьма полезно, если матрица возникла при решении задачи методом наименьших квадратов при условии, что в нашем распоряжении имеется достаточное число десятичных знаков, чтобы компенсировать накопление ошибок округления. Если заданная матрица A не слишком косоугольна, то нам удастся триангуляризация без последовательного внесения поправок. После этого мы обратим матрицу P и получим A^{-1} в виде

$$A^{-1} = \tilde{Q}Q. \quad (2-19.5)$$

Если A не симметрична, то предварительной симметризации по формуле (1) — умножением слева на $\tilde{A}$ — нельзя избежать. Однако и тогда метод триангуляризации значительно проще более громоздкого метода § 18. Однако последний всегда надежен и всегда окажется предпочтительным, если заданная матрица очень косоугольна или если порядок ее столь высок, что дополнительное количество десятичных знаков все же не может компенсировать накопление ошибок округления.

20. Обращение комплексной матрицы

Иногда приходится решать системы линейных уравнений с комплексными коэффициентами, в связи с чем возникает вопрос, как обратить матрицу, элементами которой служат комплексные числа. Принципиально эта задача не требует особого изучения, так как

комплексные числа подчиняются всем правилам обыкновенной алгебры, и все операции над комплексными числами могут производиться так же, как и над вещественными числами. Однако численные операции с комплексными числами часто приводят к ошибкам, и мы обычно предпочитаем переводить все расчеты в область вещественных чисел. Это можно сделать за счет увеличения порядка матриц с n до $2n$. Работа, связанная с обращением такой матрицы, увеличивается почти в 8 раз. Однако, в силу законов алгебры комплексных чисел, этот коэффициент увеличения можно свести к 4; это соответствует тому факту, что для перемножения двух комплексных чисел приходится выполнить четыре умножения вещественных чисел. Все же переход от комплексных матриц к вещественным целесообразен, так как мы много выигрываем в простоте проводимых операций.

Пусть C — комплексная матрица; она может быть разбита на вещественную и мнимую части:

$$C = A + iB. \quad (2-20.1)$$

Рассмотрим решение уравнения

$$Cz = c. \quad (2-20.2)$$

Как вектор z , так и вектор c комплексны. Запишем их в форме

$$z = x + iy, \quad c = a + ib. \quad (2-20.3)$$

Тогда матричное уравнение

$$(A + iB)(x + iy) = a + ib \quad (2-20.4)$$

распадается на два уравнения с вещественными элементами:

$$Ax - By = a, \quad Bx + Ay = b. \quad (2-20.5)$$

Составляющие векторов x и y можно рассматривать как $2n$ составляющих одного вектора. То же можно сделать с двумя векторами a и b :

$$\begin{aligned} \bar{x} &= (x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_n), \\ \bar{a} &= (a_1, a_2, \dots, a_n, b_1, b_2, \dots, b_n). \end{aligned}$$

Тогда линейная система (5) может быть записана в форме

$$\bar{C}\bar{x} = \bar{a}, \quad (2-20.6)$$

где матрица $2n$ -го порядка $\bar{C}$ определяется формулой

$$\bar{C} = \begin{bmatrix} A & -B \\ B & A \end{bmatrix}. \quad (2-20.7)$$

Мы можем, однако, избежать обращения этой «большой матрицы»,

если предварительно выполним обращение матрицы A , т. е. найдем A^{-1} . Мы тогда умножаем слева равенство (5) на A^{-1} и, таким образом, обособляем x :

$$x - A^{-1}By = A^{-1}a. \quad (2-20.8)$$

Таким образом x может быть выражено через y :

$$x = A^{-1}By + A^{-1}a. \quad (2-20.9)$$

Подставляя это значение во второе равенство (5), получаем

$$(A + BA^{-1}B)y = b - BA^{-1}a. \quad (2-20.10)$$

Теперь мы обратим матрицу n -го порядка

$$\bar{A} = A + BA^{-1}B.$$

Пусть обращение этой матрицы будет A_1 :

$$A_1 = (A + BA^{-1}B)^{-1}.$$

После надлежащих упрощений окончательный результат нашей схемы обращения может быть выражен следующим образом. Обращение C^{-1} комплексной матрицы $C = A + iB$ записывается в виде

$$C^{-1} = A_1 - iB_1, \quad (2-20.11)$$

где

$$\begin{aligned} A_1 &= (A + BA^{-1}B)^{-1}, \\ B_1 &= A_1BA^{-1} = A^{-1}BA_1. \end{aligned} \quad (2-20.12)$$

21. Решение кодиagonalных систем

Во многих проблемах анализа встречается особый тип системы линейных алгебраических уравнений, которая возникает в результате замены заданного линейного дифференциального уравнения разностным уравнением. Матрица такой линейной системы содержит много нулей. Ненулевые элементы матрицы расположены только на ее главной диагонали и на нескольких «кодиagonalях» (прямых, параллельных главной диагонали и расположенных по обе стороны от нее). Следовательно, неисчезающие элементы матрицы содержатся в относительно узкой полосе около главной диагонали. Такие системы могут быть решены без проведения всех выкладок, необходимых при общем методе обращения. Если число кодиagonalей, занимаемых элементами, невелико, то первоначальная система с большим числом уравнений может быть сведена к системе, содержащей лишь столько неизвестных, сколько имеется «ненулевых» кодиagonalей.

Прежде чем обратиться к изложению соответствующих алгебраических деталей, поясним прием обращения кодиagonalной матрицы графически. На рис. 13 схематически изображена система с тремя кодиagonalями. Кодиagonalи, расположенные *ниже* главной диагонали,

не обозначены на рисунке намеренно, так как для общего приема безразлично, будет ли часть матрицы ниже главной диагонали заполнена вся элементами или же она состоит из нескольких кодиагоналей, а остальные элементы равны нулю. Выделим из первоначальной матрицы, содержащей n строк и n столбцов, матрицу, обведенную жирной чертой; она содержит $n - m$ строк и $n - m$ столбцов и имеет чисто *треугольную форму*. Так как мы знаем, как обращать треугольную матрицу, то система линейных уравнений с матрицей $\bar{A}$ может быть легко решена.

Первоначальное уравнение имеет вид

$$Ax = b. \quad (2-21.1)$$

Описанную выше треугольную матрицу, получаемую опусканием в матрице A первых m столбцов и последних m строк, обозначим через $\bar{A}$. Для вектора, который получается, если в векторе $x = (x_1, x_2, \dots, x_n)$ опустить первые m составляющих, введем обозначение $\bar{x}$:

$$\bar{x} = (x_{m+1}, x_{m+2}, \dots, x_n). \quad (2-21.2)$$

Опускание последних m уравнений допустимо, так как это означает только то, что мы временно не пользуемся той информацией, которая

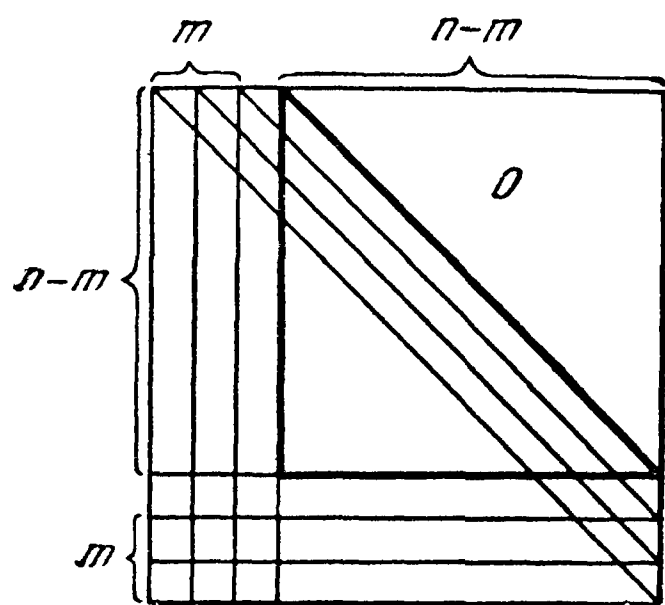


Рис. 13.

содержится в этих последних m уравнениях. Позднее мы удовлетворим и этим уравнениям. Перенесем в правую сторону уравнений элементы первых m столбцов матрицы. Если столбцы первоначальной матрицы A обозначены через $u_1, u_2, \dots, u_n$, то $\bar{u}_1, \bar{u}_2, \dots, \bar{u}_m$ будут обозначать соответствующие столбцы, в которых опущены последние m элементов. Аналогично величину $\bar{b}$ следует понимать как вектор, полученный из b , если опустить последние m его координат. Следовательно, заданная система, если не принимать

во внимание последние m уравнений, может быть записана в виде

$$\bar{A}\bar{x} = \bar{b} - (x_1\bar{u}_1 + x_2\bar{u}_2 + \dots + x_m\bar{u}_m) \quad (2-21.3)$$

(на рис. 13 принято $m=3$).

Но $\bar{A}$ есть треугольная матрица P , для которой обратная матрица Q может быть построена с помощью алгоритма, изложенного в § 17. Мы, таким образом, получаем

$$\bar{x} = Q\bar{b} - x_1Q\bar{u}_1 - x_2Q\bar{u}_2 - \dots - x_mQ\bar{u}_m. \quad (2-21.4)$$

Это равенство выражает $x_{m+1}, x_{m+2}, \dots, x_n$ через первые m неизвестных $x_1, x_2, \dots, x_m$.

Теперь мы обращаемся к решению последних m уравнений. Это уже нетрудно, так как все x_k для $k > m$ выражены через первые m неизвестных. Следовательно, мы имеем систему m линейных уравнений с m неизвестными, решение которой не представит особых трудностей, так как m — небольшое число. Первоначальная обширная система n уравнений *) сведена к гораздо меньшей системе только из m уравнений. Наконец после решения этой последней системы мы подставляем значения $x_1, x_2, \dots, x_m$ в равенство (4) и получаем, таким образом, полное решение нашей исходной задачи.

Обращение треугольной матрицы P требует, чтобы ни один из диагональных элементов не был нулем. Если некоторые из таких элементов равны нулю, то можно все же применить рассматриваемый метод со следующим видоизменением. Замещаем нули единицей и добавляем в правой части уравнений компенсирующие члены. Допустим, к примеру, что в случае задачи, изображенной на нашем рисунке, третья кодиагональ содержит нули в четвертом и пятом уравнении. Эти нули замещаются тогда единицами и соответственно добавляются x_7 в правой части четвертого и x_8 в правой части пятого уравнения. После обращения системы мы получаем как будто избыточное число неизвестных, так как x_i выражаются не только через x_1, x_2, x_3 , но и через x_7 и x_8 , тогда как последние три уравнения не могут определить больше, чем три неизвестных. В действительности, однако, четвертое и пятое уравнения дают нам два дополнительных условия, которые еще не удовлетворены. Таким образом, мы получаем $3 + 2 = 5$ уравнений для $3 + 2 = 5$ неизвестных, которые могут быть решены при условии, что система неособая.

22. Обращение матриц путем подразделения на блоки

Непосредственное обращение большой матрицы **) часто бывает нецелесообразно ввиду возможного накопления ошибок округления. Непосредственного обращения большой матрицы можно избежать путем подразделения ее на меньшие составляющие (блоки). Рассмотрим систему двадцати уравнений с двадцатью неизвестными, требующую обращения матрицы 20-го порядка. Может случиться, что у нас уже запрограммировано обращение матрицы десятого порядка. В этом случае целесообразно воспользоваться изложенным ниже методом, сводящим обращение матрицы двадцатого порядка к обращению нескольких матриц десятого порядка. В общих чертах этот метод

*) Словами «large set of equations» автор пользуется для обозначения системы уравнений с большим числом неизвестных. Для простоты мы будем пользоваться словами «обширная система».

**) То есть матрицы с большим числом элементов.

приведения большой матрицы к комбинации меньших матриц сводится к следующему. Рассмотрим матричное уравнение

$$Az = c. \quad (2-22.1)$$

Разобьем заданную большую матрицу A на четыре меньшие части, а именно две квадратные матрицы A_1 и B_2 и две прямоугольные матрицы B_1 и A_2 (рис. 14). Мы разбиваем на меньшие составляющие также неизвестный вектор z и заданную правую часть c :

$$\begin{aligned} z &= (x_1, x_2, \dots, x_{m_1}), (y_1, y_2, \dots, y_{m_2}); & x &= (x_1, x_2, \dots, x_{m_1}), \\ & & y &= (y_1, y_2, \dots, y_{m_2}), \\ c &= (a_1, a_2, \dots, a_{m_1}), (b_1, b_2, \dots, b_{m_2}); & a &= (a_1, a_2, \dots, a_{m_1}), \\ & & b &= (b_1, b_2, \dots, b_{m_2}). \end{aligned}$$

Теперь мы можем записать m_1 первых уравнений заданной системы так:

$$A_1x + B_1y = a,$$

а последние $m_2 = n - m_1$ уравнений в виде

$$A_2x + B_2y = b.$$

Таким образом, вся система уравнений оказывается эквивалентной системе двух векторных уравнений

$$\left. \begin{aligned} A_1x + B_1y &= a, \\ A_2x + B_2y &= b. \end{aligned} \right\} \quad (2-22.2)$$

Предположим, что мы можем обратить матрицу A_1 и получить A_1^{-1} .

Умножаем слева первое уравнение (2) на A_1^{-1} и уединяем таким образом x :

$$x = A_1^{-1}a - A_1^{-1}B_1y. \quad (2-22.3)$$

Подстановка этого значения во второе уравнение дает

$$(B_2 - A_2A_1^{-1}B_1)y = b - A_2A_1^{-1}a. \quad (2-22.4)$$

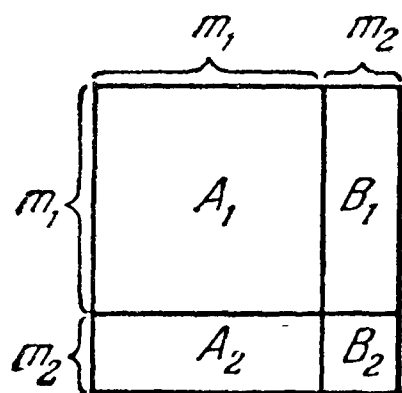


Рис. 14.

Мы здесь встречаемся с произведением матриц, не имеющих квадратной формы, т. е. число строк и число столбцов в них не одно и то же. Но принятое нами правило умножения двух матриц не предполагает обязательного равенства числа строк и столбцов в каждой матрице. Так как строки первой матрицы умножаются на столбцы второй матрицы, то необходимо только, чтобы длина *строк* первого множителя была равна длине *столбцов* второго множителя. Но это есть и единственное

условие, и в общем случае мы можем умножать матрицу с n строками и l столбцами на матрицу с l строками и m столбцами. Произведение будет матрицей с n строками и m столбцами, как это и показано на рис. 15.

Но A_2 есть матрица с m_2 строками и m_1 столбцами, A_1^{-1} — матрица с m_1 строками и m_1 столбцами. Следовательно, $(m_2, m_1) \times (m_1, m_1) = (m_2, m_1)$, и поэтому произведение $A_2 A_1^{-1}$ есть матрица с m_2 строками и m_1 столбцами. Умножая на B_1 , которая представляет собой матрицу с m_1 строками и m_2 столбцами, получаем

$(m_2, m_1) \cdot (m_1, m_2) = (m_2, m_2)$, следовательно, произведение $A_2 A_1^{-1} B_1$ есть квадратная матрица с m_2 строками и m_2 столбцами, которую можно сложить с B_2 .

Теперь мы обратим матрицу

$$\bar{B}_2 = B_2 - A_2 A_1^{-1} B_1 \quad (2-22.5)$$

и, таким образом, решим уравнение (4):

$$y = \bar{B}_2^{-1} b - \bar{B}_2^{-1} A_2 A_1^{-1} a. \quad (2-22.6)$$

Введем следующие обозначения:

$$C_2 = \bar{B}_2^{-1} A_2 A_1^{-1}; \quad C_1 = A_1^{-1} B_2 \bar{B}_2^{-1}.$$

Тогда

$$x = (A_1^{-1} + A_1^{-1} B_1 C_2) a - C_1 b; \quad y = -C_2 a + \bar{B}_2^{-1} b.$$

Вид матрицы этой системы, обратной системе (1), схематически изображен на рис. 16.

Этот метод обращения большой матрицы путем разбиения на меньшие матрицы можно рассматривать как обобщение гауссова ме-

тода исключения. Там мы рассматриваем каждое уравнение в отдельности и исключаем в нем одно неизвестное. Процесс должен быть поэтому повторен n раз. Объединяя некоторое число уравнений в одну систему и обращая соответствующую матрицу, мы можем сразу исключить целую группу неизвестных и, таким образом, существенно уменьшить систему. Например, если у нас запрограммировано обращение матрицы пятого порядка, то система двадцати уравнений с двадцатью неизвестными может быть све-

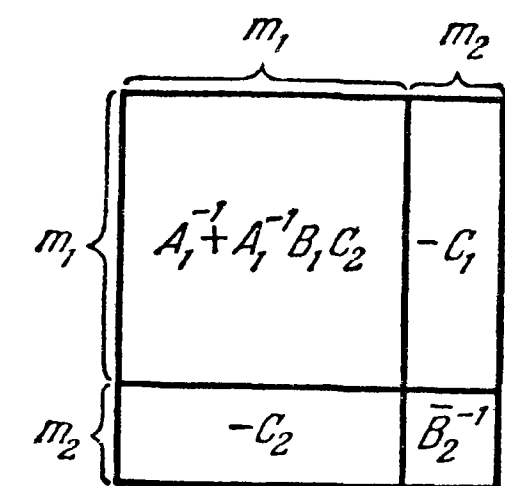


Рис. 16.

дена к системе пятнадцати уравнений с пятнадцатью неизвестными путем исключения пяти неизвестных. Затем мы снова исключаем пять неизвест-

ных и приходим к системе десяти уравнений. Следующее исключение приведет к системе пяти уравнений с пятью неизвестными, которая может быть решена непосредственно. Процесс исключения придется повторить 4 раза, тогда как первоначальный метод Гаусса потребовал бы 20 последовательных преобразований. Метод подразделения на блоки имеет еще то преимущество, что накопление ошибок округления значительно замедлится, если каждое отдельное обращение выполнять с большой точностью, применяя проверку и последующую корректировку (ср. следующий параграф).

23. Метод возмущений

Термин «возмущение» заимствован математикой из астрономии. После того как Ньютон открыл закон тяготения и заложил основания для точных математических расчетов, орбиты планет могли быть предвычислены с высокой степенью точности. Однако, предсказанное движение не согласовывалось с достаточной точностью с фактически наблюдаемым *). Это расхождение вызвано тем фактом, что планеты притягиваются не только Солнцем, но также и взаимно притягивают друг друга. Такое взаимное влияние значительно меньше влияния Солнца, так как массы планет гораздо меньше массы Солнца. Тем не менее это влияние не пренебрежимо мало, и его надо добавить к воздействию Солнца в качестве *поправки*. Таким образом, движение каждой планеты «возмущается» добавочным влиянием других планет, и это «возмущение» может быть математически вычислено.

Если бы массы планет были того же порядка, что и масса Солнца, то математическая задача определения планетных орбит была бы практически неразрешима **). Решение становится возможным ввиду того, что мы можем получить хорошее приближение, пренебрегая сначала взаимным притяжением планет, а затем учесть его в виде небольших поправок. Открытие планеты Нептун; основанное на сложных вычислениях возмущений планеты Уран, выдающимся французским астрономом Леверье и, в наше время, открытие планеты Плутон, основанное на исследованиях возмущений орбиты Нептуна, являются замечательными примерами применения остроумного метода возмущений.

В матричной алгебре методы возмущений могут быть использованы, как мощное средство противодействия вредному влиянию ошибок округления. Мы опишем здесь методы для решения линейных уравнений и обращения матриц.

*) Под предсказанным движением здесь подразумевается эллиптическое движение, определяемое притяжением одного только Солнца.

**) Такого рода задача все же может быть разрешена в настоящее время, но только с помощью электронных вычислительных машин, с затратой громадного машинного времени.

Мы получали бы точные результаты, если бы не ошибки округления, которые мешают точности наших вычислений и вызывают небольшие ошибки, накапливающимся влиянием которых уже нельзя пренебречь. Мы всегда можем проконтролировать наши результаты, подставив их в исходные уравнения и проверив, с какой точностью совпадают левые и правые части уравнений. Так как нас не интересует абсолютная точность, то иногда полученный результат мы можем считать достаточно точным. Но столь же возможно, что уравнения не удовлетворяются с желаемой точностью. Должны ли мы в этом случае отбросить все наши расчеты и начать их снова, удвоив или утроив их точность?

На самом деле это редко является необходимым. Даже неточный результат может оказаться ценным, так как с его помощью на основе метода возмущений можно вычислить более точный результат, не повторяя с самого начала все расчеты. Все методы возмущений характеризуются определенным «параметром возмущения» ϵ , который должен быть достаточно мал. Мы разлагаем точное решение в ряд по степеням ϵ и получаем, таким образом, бесконечный ряд, но фактически нам понадобятся лишь несколько членов этого ряда. Если величина ϵ имеет порядок 10^{-2} , то вторая степень ϵ будет порядка 10^{-4} , а третья степень — порядка 10^{-6} . Поэтому вряд ли придется идти в нашем разложении дальше члена третьей степени. Как правило, наш ряд будет всегда достаточно быстро сходиться, если мы начнем с первого приближения, не слишком далекого от верного решения. Тогда можно будет получить поправки на основе схемы последовательных итераций, начиная с первого грубого результата. Этот первый результат может быть сам по себе недостаточно точным, но должен быть все же достаточно близок к истинному, чтобы служить основой для метода возмущений.

Мы применим метод возмущений к трем важным случаям. Рассмотрим сначала задачу обращения матрицы. Пусть B — точная обратная матрица для заданной матрицы A , т. е.

$$BA = I. \quad (2-23.1)$$

В действительности же мы получили приближенную обратную матрицу $\bar{B}$, которая при умножении на A дает лишь приближенно матрицу I . Элементы главной диагонали матрицы $\bar{B}A$ не равны в точности единице, а элементы недиагональные не в точности равны нулю. Вообще, мы можем положить

$$\bar{B}A = I + \epsilon C. \quad (2-23.2)$$

Величина ϵ может быть нормирована требованием, чтобы наибольший по абсолютной величине элемент матрицы C был равен 1. Чем меньше будет ϵ , тем скорее будет сходиться весь дальнейший процесс.

Предположим теперь, что матрица B разложена в бесконечный ряд по степеням ε :

$$B = \bar{B} + \varepsilon B_1 + \varepsilon^2 B_2 + \varepsilon^3 B_3 + \dots \quad (2-23.3)$$

Так как B есть в точности обратная матрица для A , то должно соблюдаться равенство:

$$(\bar{B} + \varepsilon B_1 + \varepsilon^2 B_2 + \varepsilon^3 B_3 + \dots) A = I. \quad (2-23.4)$$

Переносим первый член в правую сторону, используя (2) и сокращая на ε , получаем

$$(B_1 + \varepsilon B_2 + \varepsilon^2 B_3 + \dots) A = -C.$$

Умножаем справа обе части этого равенства на соответствующие части равенства (3). В левой стороне второй множитель отпадает, так как $AB = I$, и мы получаем

$$B_1 + \varepsilon B_2 + \varepsilon^2 B_3 + \dots = -C(\bar{B} + \varepsilon B_1 + \varepsilon^2 B_2 + \dots). \quad (2-23.5)$$

Приравнявая коэффициенты при одинаковых степенях ε в обеих частях этого равенства, получаем следующую последовательность уравнений, которая позволяет нам определить $B_1, B_2, B_3, \dots$ рекуррентно:

$$\left. \begin{aligned} B_1 &= -C\bar{B}, \\ B_2 &= -CB_1, \\ B_3 &= -CB_2, \\ &\dots \end{aligned} \right\} \quad (2-23.6)$$

Небольшое число шагов обычно дает достаточную точность. Практически большое число этих последовательных матричных умножений не приведет к желаемой цели, так как при этом появятся ошибки округления, которые в конце концов станут большими, чем ошибка основной операции. Из окончательных формул мы можем исключить параметр ε , положив

$$\bar{B}A = I + C, \quad (2-23.7)$$

т. е. заменяя в (2) εC на C . Тогда

$$B = \bar{B} + B_1 + B_2 + B_3 + \dots,$$

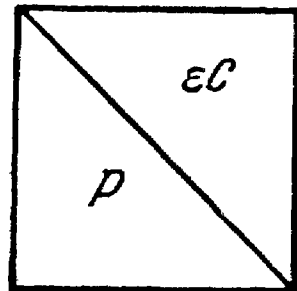
где

$$\left. \begin{aligned} B_1 &= -C\bar{B}, \\ B_2 &= -CB_1, \\ B_3 &= -CB_2, \\ &\dots \end{aligned} \right\} \quad (2-23.8)$$

На каком бы шагу мы ни остановились, мы можем взять результирующую матрицу B и, умножив ее на A , установить, насколько близко полученное произведение к единичной матрице I . Если предпосылки

для успешного применения метода возмущений были выполнены, то остаточные ошибки после небольшого числа приближений окажутся пренебрежимо малыми.

В качестве второго примера приложения метода возмущений рассмотрим случай матрицы *почти треугольной*. В ней элементы, расположенные на главной диагонали и ниже ее, образуют треугольную матрицу P . Элементы над диагональю не равны нулю, но малы по сравнению с диагональными элементами матрицы P . Все эти элементы мы объединяем в матрицу εC (рис. 17) и полагаем



$$A = P + \varepsilon C. \quad (2-23.9) \quad \text{Рис. 17.}$$

Обращение треугольной матрицы P не представит теперь труда, и мы, таким образом, получим новую треугольную матрицу Q :

$$QP = I. \quad (2-23.10)$$

Мы можем принять, что это равенство удовлетворяется с ошибкой, которая мала сравнительно с εC .

Разлагая теперь точную обратную матрицу $A^{-1} = B$ в бесконечный ряд по степеням ε , получим:

$$B = Q + \varepsilon B_1 + \varepsilon^2 B_2 + \varepsilon^3 B_3 + \dots \quad (2-23.11)$$

Тогда, в силу определения обратной матрицы,

$$\left. \begin{aligned} (Q + \varepsilon B_1 + \varepsilon^2 B_2 + \varepsilon^3 B_3 + \dots)(P + \varepsilon C) &= I, \\ QP + \varepsilon B_1 P + \varepsilon^2 B_2 P + \varepsilon^3 B_3 P + \dots & \\ \dots + \varepsilon QC + \varepsilon^2 B_1 C + \varepsilon^3 B_2 C + \dots &= I. \end{aligned} \right\} \quad (2-23.12)$$

Так как, однако, $QP = I$, то, приравнявая нулю каждый коэффициент при степенях ε в этом разложении, получаем

$$\begin{aligned} B_1 P &= -QC, \\ B_2 P &= -B_1 C, \\ B_3 P &= -B_2 C \end{aligned}$$

и после умножения этих равенств справа на Q :

$$\left. \begin{aligned} B_1 &= -QCQ, \\ B_2 &= -B_1 CQ, \\ B_3 &= -B_2 CQ, \\ &\dots \end{aligned} \right\} \quad (2-23.13)$$

Из наших результатов мы снова можем исключить параметр разложения ε :

$$\left. \begin{aligned} A &= P + C, \\ B &= Q + B_1 + B_2 + B_3 + \dots, \\ B_1 &= -QCQ, \\ B_2 &= -B_1CQ, \\ B_3 &= -B_2CQ, \\ &\dots \end{aligned} \right\} \quad (2-23.14)$$

Здесь опять можно остановиться на каком-нибудь B_k , где k выбирается надлежащим образом (оно редко будет превышать 4), и снова проверить точность нашего результата путем образования произведения полученной матрицы B на A . Если условия для успешного применения метода возмущений были подходящими, то отклонение этого произведения от единичной матрицы будет пренебрежимо малым.

В качестве третьего примера приложения метода возмущений рассмотрим решение системы линейных уравнений:

$$Ax = b. \quad (2-23.15)$$

Предположим, что мы каким-либо образом получили решение этой задачи, не обращая матрицу A . Однако, подстановка этого решения в (15) дает b не точно, а лишь приближенно.

Найденное приближенное решение обозначим через $\bar{x}$. Разность

$$b - A\bar{x} = b_1 \quad (2-23.16)$$

назовем «остаточным вектором» нашего решения. Наша задача теперь сведена к решению новой системы

$$Ax_1 = b_1, \quad x_1 = x - \bar{x}. \quad (2-23.17)$$

И для этого уравнения мы опять найдем не точное решение, а лишь приближенное $\bar{x}_1$, для которого опять находится новый остаточный вектор b_2 :

$$b_1 - A\bar{x}_1 = b_2, \quad (2-23.18)$$

что приводит нас к новому уравнению

$$Ax_2 = b_2, \quad (2-23.19)$$

$$x = \bar{x} + \bar{x}_1 + x_2. \quad (2-23.20)$$

Этот процесс может быть продолжен до тех пор, пока мы не придем к остаточному вектору b_k , который столь мал, что соответствующим x_k можно пренебречь. Тогда x оказывается равным сумме частичных решений $\bar{x} + \bar{x}_1 + \bar{x}_2 + \dots + \bar{x}_{k-1}$. Этот метод последовательных приближений применим всегда, когда пользуются машиной

непрерывного действия для решения системы линейных уравнений. Такая машина (механического или электрического типа) редко дает решение с точностью, большей чем $1^0/0$. Это значит, что исходная правая часть уравнений может быть уменьшена на две значащие цифры при подстановке найденного решения в заданную систему.

Получив новое решение и подставив его в остаточную систему, мы снова уменьшим правую часть на две значащие цифры. Следовательно, в четыре последовательных приема мы получим 8 значащих цифр. Таким образом, первоначальное решение умеренной точности уточнено до решения большой точности. Этот же принцип применим всегда, когда мы обладаем каким-нибудь методом (исключения или итерации), с помощью которого можно получить умеренно точное решение заданной системы линейных уравнений.

24. Совместимость линейных уравнений

Чисто математическое решение системы линейных уравнений часто вуалирует опасности, которые возникают при решении обширных систем линейных уравнений. Мы склонны использовать большие возможности современных электронных вычислительных машин для решения обширных линейных систем, не учитывая, что точное математическое решение, полученное таким образом, может не иметь никакого физического смысла. Поэтому следует рассмотреть вопрос о физическом значении математически верного решения и обсудить проблему «помех». «Помехи», которые здесь имеются в виду, касаются не «арифметических помех», получающихся в результате ошибок округления при вычислениях, а «физических помех», обусловленных неточностью наших измерений.

Следующий пример очень удобен для упрощенной характеристики трудностей, возникающих при решении линейных систем, имеющих весьма «косоугольный» характер. Рассмотрим два уравнения

$$x + y = 2,00001, \quad x + 1,00001y = 2,00002. \quad (2-24.1)$$

Эта система имеет решение

$$x = 1,00001, \quad y = 1.$$

С чисто математической точки зрения мы имеем два уравнения с двумя неизвестными, и система не особая, так как определитель системы отличен от нуля. Следовательно, система допускает единственное решение, и, найдя его, мы нашей цели достигли.

С физической точки зрения дело обстоит совсем иначе. Правая часть уравнений обычно возникает в результате физических измерений, а эти измерения имеют ограниченную точность. Поэтому правые части вышеуказанной системы могли быть заданы с точностью не до пяти десятичных знаков, а только до двух десятичных знаков. Но тогда сразу видно, что наша система не может быть решена

относительно x и y . Причина состоит в том, что неизвестные x и y нашей задачи входят в уравнения практически лишь своей *суммой*. Значение суммы $x + y$ легко получить из заданных уравнений. Но для разделения x и y нам нужна также разность $x - y$. Но, если положить

$$\frac{1}{2}(x + y) = \xi, \quad \frac{1}{2}(x - y) = \eta,$$

то наши уравнения преобразуются к виду

$$2\xi = 2,00001, \quad 2,00001\xi - 0,00001\eta = 2,00002. \quad (2-24.2)$$

Таким образом, величина η лишь *очень слабо* представлена в нашей системе, и следовательно, требуется *исключительная точность* при ее вычислении. Для получения η надо делить на 10^{-5} , т. е. умножать на 10^5 ; поэтому правая часть должна быть известна с такой исключительной точностью, какая обычно невозможна по физическим причинам. Если бы, например, физические помехи наблюдений вызвали бы ошибку во втором уравнении на 0,001, то ξ практически осталось бы равным 1, тогда как η стало бы равным 100, что привело бы к совершенно ошибочному решению $x = 101$, $y = -99$. Необходимо уяснить себе, что при данных физических условиях мы можем из наших уравнений определить с достаточной точностью сумму $x + y$, но о раздельном определении x и y не может быть речи.

В нашем простом примере мы могли проследить каждую деталь процесса решения и явно показать неудовлетворительный характер заданной почти особой системы. Однако в более сложных случаях мы часто принимаем за истину результаты, полученные при решении весьма косоугольных систем, не учитывая того обстоятельства, что, в силу физических помех, математическое решение может оказаться весьма далеким от истинных значений тех величин, которые мы как будто определили в своем решении.

Чтобы установить надлежащий критерий для оценки достоверности решений весьма косоугольных систем, следует сначала развить математическую теорию совместности линейных систем, а затем соответственно видоизменить ее для применения к вопросу о физической допустимости данной системы линейных уравнений.

Математическая задача совместности линейных систем определяется следующими соображениями. Если задача содержит n неизвестных, то мы прежде всего постараемся получить n уравнений для определения этих неизвестных. Если число уравнений меньше чем n , то мы наперед знаем, что имеющаяся информация недостаточна для однозначного определения всех неизвестных. Если число неизвестных больше чем n , то наша информация избыточна и, вообще говоря, приведет к противоречиям. Мы, следовательно, исключаем недостаточно определенные системы, так как они содержат слишком мало информации и не могут привести к однозначному решению задачи,

и исключаем переопределенные системы, так как они содержат слишком много информации и не могут принести пользы.

Следует, однако, ясно сознавать, что совместность заданной системы уравнений не имеет никакого отношения к недостаточно определенному, переопределенному или точно определенному («сбалансированному») характеру задачи. Следующая система двух уравнений недостаточно определена, так как она состоит из двух уравнений с пятью неизвестными:

$$\left. \begin{aligned} x_1 + x_2 + x_3 + x_4 + x_5 &= 3, \\ 2x_1 + 2x_2 + 2x_3 + 2x_4 + 2x_5 &= 8. \end{aligned} \right\} \quad (2-24.3)$$

Однако эти два уравнения внутренне противоречивы и не могут быть удовлетворены ни при каких значениях пяти неизвестных. С другой стороны, следующие пять уравнений содержат только два неизвестных и, таким образом, система переопределена:

$$\left. \begin{aligned} x_1 + x_2 &= 0, \\ 2x_1 + 3x_2 &= -1, \\ 3x_1 + 2x_2 &= 1, \\ x_1 - x_2 &= 2, \\ 3x_1 + 5x_2 &= -2. \end{aligned} \right\} \quad (2-24.4)$$

Тем не менее эти пять уравнений не противоречивы и имеют решение

$$x_1 = 1, \quad x_2 = -1.$$

Возникает вопрос, существует ли какой-либо систематический метод определения совместности или несовместности заданной системы уравнений. Такой метод, действительно, существует и может быть описан следующим образом. Рассмотрим линейное уравнение

$$Ax = b, \quad (2-24.5)$$

а вместе с ним и сопряженное ему однородное уравнение

$$\tilde{A}y = 0. \quad (2-24.6)$$

Пусть x_0 — какое-либо решение уравнения (5), а y_0 — какое-либо решение уравнения (6), т. е. пусть выполняются равенства

$$Ax_0 = b \quad \text{и} \quad \tilde{A}y_0 = 0.$$

Умножив первое из них скалярно на вектор y_0 , а второе на вектор x , получим

$$y_0 \cdot Ax_0 = y_0 \cdot b, \quad x_0 \cdot \tilde{A}y_0 = 0. \quad (2-24.7)$$

Согласно основному правилу транспонирования (8.23), имеем

$$y_0 \cdot Ax_0 = x_0 \cdot \tilde{A}y_0. \quad (2-24.8)$$

Поэтому из (7) следует

$$y_0 \cdot b = 0. \quad (2-24.9)$$

Мы пришли, таким образом, к следующему результату: для того чтобы уравнение (5) имело решение, необходимо, чтобы расположенный в его правой части заданный вектор b был перпендикулярен к каждому из решений однородного уравнения (6).

Можно показать, что это условие является не только необходимым, но и достаточным; оно представляет собой поэтому некоторый общий принцип, позволяющий разрешить все вопросы о совместности заданной системы линейных уравнений.

Этот общий принцип в каждом отдельном случае по-разному применяется. Во-первых, может случиться, что сопряженное однородное уравнение $\tilde{A}y = 0$ *совсем не имеет решений* кроме тождественно равного нулю. В этом случае мы не получаем никакого условия совместности, что означает, что заданная система (5) совместна при *любых* значениях правой части. Во-вторых, возможно, что сопряженное однородное уравнение $\tilde{A}y = 0$ имеет одно и только одно решение (не считая тривиального решения $y = 0$), определенное с точностью до произвольного множителя ввиду однородности уравнения. В этом случае вектор правой части заданного уравнения должен удовлетворять одному условию совместности: быть ортогональным решению сопряженного однородного уравнения. В-третьих, может случиться, что сопряженное однородное уравнение $\tilde{A}y = 0$ имеет несколько независимых решений. В этом случае вектор правой части заданного уравнения должен быть ортогонален *каждому* из этих независимых решений.

Вопросы об определенности, переопределенности или недостаточной определенности системы не находят своего отражения в этом общем принципе, за исключением того факта, что в случае переопределенной системы сопряженная система *всегда* имеет нетривиальные решения и, следовательно, правая часть заданного уравнения всегда должна удовлетворять одному или нескольким условиям.

Примеры. 1. Рассмотрим линейную систему

$$\left. \begin{aligned} x_1 + x_2 + x_3 &= 1, \\ 2x_1 + 2x_2 + 3x_3 &= 3. \end{aligned} \right\}$$

Здесь матрица A есть

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 2 & 3 \end{bmatrix}.$$

Сопряженная матрица $\tilde{A}$:

$$\tilde{A} = \begin{bmatrix} 1 & 2 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}.$$

Ей соответствует сопряженная система уравнений

$$\left. \begin{aligned} y_1 + 2y_2 &= 0, \\ y_1 + 2y_2 &= 0, \\ y_1 + 3y_2 &= 0. \end{aligned} \right\}$$

Эти уравнения не имеют нетривиального решения и, следовательно, заданные уравнения совместны, независимо от того, каковы их правые части.

2. Теперь рассмотрим систему

$$\left. \begin{aligned} x_1 + x_2 + x_3 &= 1, \\ 2x_1 + 2x_2 + 2x_3 &= 1. \end{aligned} \right\}$$

Сопряженной системой является

$$\left. \begin{aligned} y_1 + 2y_2 &= 0, \\ y_1 + 2y_2 &= 0, \\ y_1 + 2y_2 &= 0. \end{aligned} \right\}$$

Эта система имеет решение

$$y_1 = -2, \quad y_2 = 1.$$

Следовательно, правая часть системы должна удовлетворять условию

$$-2c_1 + c_2 = 0, \quad \text{или} \quad c_2 = 2c_1.$$

Заданные правые части уравнений не удовлетворяют этому условию, и следовательно, уравнения несовместны и, значит, неразрешимы.

3. Рассмотрим переопределенную систему

$$\left. \begin{aligned} x_1 + x_2 &= 0, \\ 2x_1 + 3x_2 &= -1, \\ 3x_1 + 2x_2 &= 1, \\ x_1 - x_2 &= 2, \\ 3x_1 + 5x_2 &= -2. \end{aligned} \right\}$$

Матрица A теперь содержит два столбца в пять строк:

$$A = \begin{bmatrix} 1 & 1 \\ 2 & 3 \\ 3 & 2 \\ 1 & -1 \\ 3 & 5 \end{bmatrix}.$$

Следовательно, сопряженная матрица имеет две строки и пять столбцов:

$$\tilde{A} = \begin{bmatrix} 1 & 2 & 3 & 1 & 3 \\ 1 & 3 & 2 & -1 & 5 \end{bmatrix}.$$

Сопряженная система содержит два уравнения с пятью неизвестными:

$$\left. \begin{aligned} y_1 + 2y_2 + 3y_3 + y_4 + 3y_5 &= 0, \\ y_1 + 3y_2 + 2y_3 - y_4 + 5y_5 &= 0. \end{aligned} \right\}$$

Эта система имеет ровно три независимых решения:

$$\begin{aligned} y_1 &= -5, & y_2 &= 1, & y_3 &= 1, & y_4 &= 0, & y_5 &= 0; \\ y_1 &= -5, & y_2 &= 2, & y_3 &= 0, & y_4 &= 1, & y_5 &= 0; \\ y_1 &= 0, & y_2 &= 9, & y_3 &= -1, & y_4 &= 0, & y_5 &= -5. \end{aligned}$$

Соответственно этому правые части уравнений должны удовлетворять трем условиям:

$$\left. \begin{aligned} -0.5 - 1 \cdot 1 + 1 \cdot 1 + 2 \cdot 0 - 2 \cdot 0 &= 0, \\ -0.5 - 1 \cdot 2 + 1 \cdot 0 + 2 \cdot 1 - 2 \cdot 0 &= 0, \\ -0 \cdot 0 - 1 \cdot 9 + 1 \cdot (-1) + 2 \cdot 0 - 2 \cdot (-5) &= 0. \end{aligned} \right\}$$

Так как эти условия действительно выполняются, то заданная переопределенная система совместна.

Из того, что некоторая система уравнений совместна, не следует обязательно, что решение системы *единственно*. Совместная система уравнений может иметь одно или много решений. В нашем втором примере мы имели недостаточно определенную несовместную систему, которая не имела решений; в первом примере — недостаточно определенную совместную систему, которая имела бесчисленное количество решений, наконец, в третьем примере — переопределенную совместную систему, которая имеет единственное решение. Вообще, можно сказать, что если удовлетворяется какое-нибудь условие совместности, то можно опустить одно из уравнений системы как излишнее. Переопределенная система, таким образом, может обратиться в определенную. Но может также случиться, что она станет недостаточно определенной. Рассмотрим, например, следующие четыре уравнения с тремя неизвестными:

$$\left. \begin{aligned} x_1 + 3x_2 - 2x_3 &= 11, \\ 2x_1 - 5x_2 + 7x_3 &= -11, \\ -x_1 + 2x_2 - 3x_3 &= 4, \\ x_1 + 2x_2 - x_3 &= 8. \end{aligned} \right\}$$

Сопряженные уравнения суть

$$\left. \begin{aligned} y_1 + 2y_2 - y_3 + y_4 &= 0, \\ 3y_1 - 5y_2 + 2y_3 + 2y_4 &= 0, \\ -2y_1 + 7y_2 - 3y_3 - y_4 &= 0. \end{aligned} \right\}$$

Они имеют два независимых решения

$$\begin{aligned} y_1 &= 1, & y_2 &= 5, & y_3 &= 11, & y_4 &= 0; \\ y_1 &= 9, & y_2 &= 1, & y_3 &= 0, & y_4 &= -11. \end{aligned}$$

Соответственно этому заданные правые части должны удовлетворять двум условиям

$$\left. \begin{aligned} 11 \cdot 1 - 11 \cdot 5 + 4 \cdot 11 + 8 \cdot 0 &= 0, \\ 11 \cdot 9 - 11 \cdot 1 + 4 \cdot 0 + 8 \cdot (-11) &= 0. \end{aligned} \right\}$$

Эти условия действительно удовлетворяются, и совместность заданной системы, таким образом, обеспечена. Соответственно общему принципу мы можем опустить столько уравнений, сколько имеется условий совместности. Это сводит число независимых уравнений к двум, и система становится недостаточно определенной, так как у нас остается лишь два уравнения с тремя неизвестными.

25. Переопределенность и принцип наименьших квадратов

Трудности, с которыми часто связано решение обширных систем линейных уравнений, можно сравнить с трудностями, с которыми встречается оратор, который старается в своей речи охватить слишком много различных тем. Вначале его речь течет очень плавно. Однако по мере того, как он находит в своем конспекте все новые темы, он все более сбивается и порой теряет нить своих мыслей. В нужный момент он забывает подробности, о которых хотел рассказать, и поэтому упускает некоторые вопросы, повторяя вместе с тем в других выражениях то, о чем уже ранее говорил. А так как он не в большом ладу с истиной, то впадает в противоречия, забывая, как он толковал предмет в своих прежних замечаниях (нарушая старый ораторский принцип: *Mendacem oportet esse memorem* — лгун должен иметь хорошую память). В последние десять минут он ничего уже не помнит, перевирает все на свете и, наконец, усаживается под гром аплодисментов.

В нашем случае лгуном с плохой памятью являются «помехи», которые мешают точности наших измерений и нарушают правильный ход событий. Так как помехи носят случайный характер, то нарушения не согласованы и происходят то в одном, то в другом направлении. Это *одна* опасность, которая встречается при учете физических явлений. *Другая* опасность состоит в том, что информация, которой мы располагаем, *недостаточна* для действительного определения всех неизвестных проблемы. Наш оратор позабыл объяснить некоторые пункты своего конспекта и заменил упущенное пересказом другими словами эпизодов, уже ранее приведенных. Аналогичное положение может встретиться и в рассматриваемых задачах: может случиться (и часто действительно случается), что утверждения нашей системы уравнений недостаточны для полного определения всех неизвестных задачи. Мы подсчитываем наши уравнения и находим, что их у нас ровно столько, сколько неизвестных. Мы считаем поэтому, что наша система определена и допускает единственное решение. Однако может оказаться, что некоторые уравнения просто повторяют другими

словами утверждения, уже ранее сделанные, не добавляя ничего существенно нового к ним. В этом случае наша система недостаточно определена и не в состоянии дать полного решения задачи.

Эти две трудности между собой связаны. Если левые части уравнений находятся в какой-либо зависимости друг от друга, то правые части должны удовлетворять определенным условиям совместности. Но эти условия могут не выполняться ввиду ошибок измерений, что делает наши уравнения несовместными в строго математическом смысле. С этой несовместностью сочетается тот факт, что наши уравнения недостаточны для определения всех неизвестных задачи, так как для достижения совместности мы должны были бы опустить некоторое число уравнений как избыточных, и в этом случае не имели бы достаточного числа уравнений для полного решения.

Таким образом, вопросы недостаточной определенности и переопределенности переплетаются между собой. Наша система фактически недостаточно определена ввиду отсутствия сведений о некоторых линейных комбинациях неизвестных, и это делает невозможным определение *всех* неизвестных систем. Соответствующее число уравнений оказывается лишним и может быть опущено. Следовательно, система n уравнений с n неизвестными, в которой опущены m линейных комбинаций неизвестных, в действительности представляет собой систему n уравнений с $n - m$ неизвестными и, таким образом, переопределена на m уравнений.

С одним из этих затруднений можно справиться, а именно с переопределенностью. Остроумный метод наименьших квадратов дает возможность выправить любую переопределенную и несовместную систему уравнений. В действительности мы даже извлекаем пользу из этого затруднения и стараемся переопределить как можно больше систему уравнений, производя некоторое число дополнительных наблюдений сверх того минимального числа, которое требуется числом неизвестных. Мы получаем при этом систему линейных уравнений, которая характеризуется матричным уравнением

$$Ax = b, \quad (2-25.1)$$

где A — не квадратная матрица, а матрица, содержащая гораздо больше строк, чем столбцов. Ввиду ошибок наших измерений уравнения оказываются математически несовместными. Это значит, что мы не можем сделать все составляющие остаточного вектора $Ax - b$ равными нулю.

Но тогда мы можем поставить вопрос о нахождении решения, которое будет «наилучшим» при данных обстоятельствах. С этой целью образуем остаточный вектор

$$Ax - b = r, \quad (2-25.2)$$

берем квадрат длины этого остаточного вектора и определяем x из условия, чтобы r^2 было наименьшим. Задача приведения к минимуму выражения $(Ax - b)^2$ всегда имеет определенное решение,

независимо от того, совместна или не совместна заданная система. Если эта система совместна, то значение x , соответствующее наименьшему квадрату, при подстановке в $Ax = b$ обратит это выражение в нуль. Если же система несовместна, то остаточный вектор $Ax - b$ не будет нулем, но найденное значение будет решением с наименьшей ошибкой в том смысле, что квадрат длины остаточного вектора будет меньше, чем при любом другом выборе x .

Таким образом, решение способом наименьших квадратов полностью избавляет нас от исследования совместности заданной системы, так как мы примирились с тем фактом, что мы не получаем точного решения нашей задачи, а только наилучшее решение, возможное при данных обстоятельствах. Если заданная система совместна, то остаточный вектор, полученный методом наименьших квадратов, само собой окажется нулем, подтверждая тем самым, что система совместна.

Решение уравнения $Ax = b$ в смысле наименьших квадратов есть нахождение вектора x , для которого скаляр $(Ax - b)(Ax - b)$ принимает наименьшее значение. В результате использования известных из анализа методов нахождения экстремальных значений скалярной функции от многих переменных $x_1, x_2, \dots, x_n$ — координат вектора x — эта задача сводится к решению уравнения

$$\tilde{A}Ax = \tilde{A}b. \quad (2-25.3)$$

Замечательное свойство этого уравнения состоит в том, что оно всегда приводит к вполне определенной системе точно такого числа уравнений, сколько у нас имеется неизвестных, независимо от того, как сильно была переопределена первоначальная система. Из рис. 18 видно, что произведение матрицы из m строк и n столбцов на матрицу из n строк и m столбцов представляет собой квадратную матрицу с m строками и m столбцами. Поэтому в окончательной системе число уравнений равняется числу неизвестных независимо от числа уравнений, которые содержала первоначальная система. Кроме того, матрица $\tilde{A}A$ всегда симметрична:

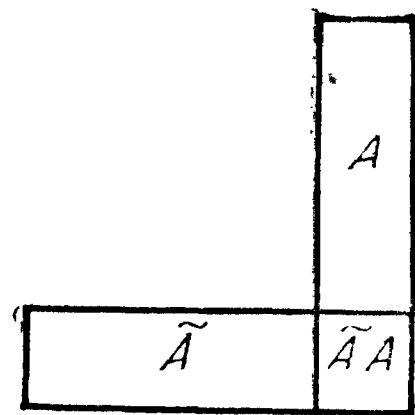


Рис. 18.

$$(\tilde{A}A) = \tilde{A}A, \quad (2-25.4)$$

и ее собственные значения не только вещественны, но и *положительны* или равны нулю.

Пример. Для переопределенной системы (24.4) мы получаем

$$\begin{bmatrix} 1 & 1 \\ 2 & 3 \\ 3 & 2 \\ 1 & -1 \\ 3 & 5 \end{bmatrix} \odot \begin{bmatrix} 1 & 1 \\ 2 & 3 \\ 3 & 2 \\ 1 & -1 \\ 3 & 5 \end{bmatrix} = \begin{bmatrix} 24 & 27 \\ 27 & 40 \end{bmatrix}.$$

Кроме того, $\tilde{A}b$ оказывается равным

$$\begin{bmatrix} 0 & -1 & 1 & 2 & -2 \\ 1 & 2 & 3 & 1 & 3 \\ 1 & 3 & 2 & -1 & 5 \end{bmatrix} = \begin{bmatrix} -3 \\ -13 \end{bmatrix}$$

Следовательно, система (3) принимает в нашем случае вид

$$\left. \begin{aligned} 24x_1 + 27x_2 &= -3, \\ 27x_1 + 40x_2 &= -13. \end{aligned} \right\} \quad (2-25.5)$$

Решением этой системы служит

$$x_1 = 1, \quad x_2 = -1.$$

Подставив это решение в первоначальную систему (24.4), находим, что все уравнения удовлетворяются. Совместность заданной системы, таким образом, доказана.

С другой стороны, в случае недостаточно определенной системы даже метод наименьших квадратов не может помочь найти единственное решение. Если в системе не хватает сведений о некоторых комбинациях неизвестных, то никакие магические заклинания не могут их вызвать. Недостаточно определенная система так и останется недостаточно определенной, даже если сформулировать ее в форме способа наименьших квадратов. Однако несовместная недостаточно определенная система преобразуется в совместную систему недостаточно определенную. Следующие четыре уравнения с тремя неизвестными представляют, на первый взгляд, переопределенную, но на самом деле недостаточно определенную и несовместную систему уравнений:

$$\left. \begin{aligned} x_1 + 3x_2 - 2x_3 &= 11, \\ 2x_1 - 5x_2 + 7x_3 &= -10, \\ -x_1 + 2x_2 - 3x_3 &= 4, \\ x_1 + 2x_2 - x_3 &= 8. \end{aligned} \right\} \quad (2-25.6)$$

Метод наименьших квадратов сводит эти уравнения к системе:

$$\left. \begin{aligned} 7x_1 - 7x_2 + 14x_3 &= -5, \\ -7x_1 + 42x_2 - 49x_3 &= 107, \\ 14x_1 - 49x_2 + 63x_3 &= -112. \end{aligned} \right\} \quad (2-25.7)$$

Матрица, транспонированная к матрице этой системы, совпадает с первоначальной; мы должны, следовательно, найти решение однородной системы

$$\left. \begin{aligned} 7y_1 - 7y_2 + 14y_3 &= 0, \\ -7y_1 + 42y_2 - 49y_3 &= 0, \\ 14y_1 - 49y_2 + 63y_3 &= 0. \end{aligned} \right\} \quad (2-25.8)$$

Оно определяется формулами

$$y_1 = 1, \quad y_2 = -1, \quad y_3 = -1. \quad (2-25.9)$$

Правые части уравнений (7) удовлетворяют условию ортогональности к этому однородному решению. Следовательно, мы можем опустить третье уравнение системы (7) и свести последнюю к системе двух уравнений

$$\left. \begin{aligned} 7x_1 - 7x_2 + 14x_3 &= -5, \\ -7x_1 + 42x_2 - 49x_3 &= 107, \end{aligned} \right\}$$

которая имеет бесчисленное количество решений. Приняв $x_3 = 0$, приходим к решению:

$$x_1 = 2,2, \quad x_2 = 2,9143, \quad x_3 = 0.$$

Если подставим это решение в левые части начальных уравнений (6), то получим справа

$$10,943, \quad -10,171, \quad 3,629, \quad 8,029,$$

вместо требуемых чисел

$$11, \quad -10, \quad 4, \quad 8,$$

которые, однако, не могут быть получены ни при каких значениях неизвестных ввиду несовместности заданных уравнений.

26. Естественная и искусственная косоугольность системы линейных уравнений

Трудность обращения матрицы весьма косоугольной системы уравнений порой без нужды увеличивается неподходящим выбором масштабов для измерения неизвестных. Мы трактовали уравнение

$$Ax = b \quad (2-26.1)$$

как задачу разложения заданного вектора b по осям заданной косоугольной системы координат, определяемым столбцами матрицы A :

$$x_1 u_1 + x_2 u_2 + \dots + x_n u_n = b. \quad (2-26.2)$$

Определитель $|A|$ линейной системы (1) имеет следующую замечательную геометрическую интерпретацию. Он представляет объем параллелепипеда, образованного косоугольными базисными векторами $u_1, u_2, \dots, u_n$ (рис. 19). Этот определитель может стать очень малым по двум причинам. Одна состоит в том, что векторы u_i образуют весьма малые углы друг с другом. Другая причина — та, что длины некоторых из этих векторов оказываются весьма малыми. Последнее обстоятельство вызвано неподходящим выбором масштабов для измерения неизвестных $x_1, x_2, \dots, x_n$ и может быть устранено путем простого изменения

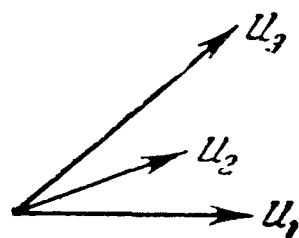


Рис. 19.

этих масштабов:

$$x_1 = \alpha_1 \bar{x}_1, \quad x_2 = \alpha_2 \bar{x}_2, \dots, x_n = \alpha_n \bar{x}_n. \quad (2-26.3)$$

Мы можем затем умножить векторы $u_1, u_2, \dots, u_n$ на произвольные постоянные множители $\alpha_1, \alpha_2, \dots, \alpha_n$ и таким путем избежать очень большого расхождения в их длине. Действительно, мы можем нормировать длину каждого столбца в точности к 1. В этом случае определитель новой системы является истинной мерой косоугольности системы. Этот определитель по абсолютной величине всегда меньше 1. Максимальное значение 1 соответствует случаю ортогональных осей. Чем меньше этот определитель, тем меньше взаимная независимость уравнений системы. В качестве нижнего предела мы можем получить определитель, равный нулю, что произойдет, если один из базисных векторов окажется линейной комбинацией остальных.

Дополнительным преимуществом такого нормирования является то, что все элементы матрицы тогда лежат между ± 1 , и следовательно, мы избегаем операций с очень большими числами.

Существует, однако, еще один процесс нормирования, на который мы должны обратить внимание. Мы видели, что при обработке системы линейных уравнений методом наименьших квадратов вычисляется сумма квадратов координат остаточного вектора, и эта сумма доводится до минимума. Но возможно, что сами уравнения недостаточно сбалансированы в процессе образования сумм квадратов. В этом случае каждое уравнение

$$a_{i1}x_1 + a_{i2}x_2 + \dots + a_{in}x_n - b_i = 0$$

перед возведением в квадрат полезно умножить на произвольный множитель β_i . Этот множитель β_i мы можем использовать для того, чтобы избежать опасности, что некоторые уравнения войдут в сумму квадратов со слишком большим или со слишком малым весом. Если вес какого-нибудь отдельного уравнения слишком мал, то это значит, что это уравнение сравнительно с другими уравнениями практически не участвует в образовании системы (25.3) или, другими словами, что одна важная часть информации теряется, и наша система практически приводится к недостаточно определенной системе. Если же вес слишком велик, то мы придаем слишком большое значение одной части информации за счет всех остальных частей, что ведет даже к худшему виду недостаточно определенной системы, практически исключая все другие ценные части информации. Наилучшее равновесие получается, если мы выбираем наши множители β_i так, чтобы сумма квадратов элементов каждой строки, не считая правой части, равнялась бы приблизительно 1. При таком нормировании мы придаем каждому уравнению одинаковую значимость и избегаем двух крайностей: излишнего или недостаточного внимания к одному какому-нибудь уравнению.

Эти два нормирования — одно относится к столбцам, другое — к строкам — не могут быть выполнены одновременно (за исключением случая симметрической матрицы). Выравнивание длины столбцов нарушит равновесие длины строк. Потребуется несколько повторных подгонок строк и столбцов, пока мы добьемся того, что суммы квадратов как строк, так и столбцов будут достаточно выравнены. Это выравнивание нет нужды проводить с большой точностью, поскольку отклонение от некоторой средней длины не будет серьезным. Двойное выравнивание гарантирует, что косоугольность системы не ухудшена искусственно неадекватным нормированием как уравнений, так и неизвестных.

27. Ортогонализация произвольной линейной системы

Мы видели, что трудность решения весьма косоугольных систем состоит в том, что некоторые линейные комбинации неизвестных входят в уравнения с исключительно малым весом. Поэтому нелегко выделить отдельно каждое неизвестное. Это потребовало бы такой точности определения правых частей уравнений, какой мы часто не располагаем.

Для общего анализа нашей системы первоначальная косоугольная система базисных векторов, представленная последовательными столбцами матрицы, неудобна, так как мы не можем эффективно оперировать с косоугольными осями. Счастливым обстоятельством является то, что при переходе к новой координатной системе, получающейся из исходной некоторым простым вращением, наши весьма косоугольные векторы, определяющие матрицу, все более и более «раскрываются». Такое вращение без каких-либо других искусственных приемов приводит к новой системе координат, в которой наши первоначально косоугольные векторы, определяющие матрицу, становятся полностью *ортогональными*. Этот прием обладает полной общностью и применим даже к особым матрицам. *Произвольную матрицу можно ортогонализировать путем простого вращения системы координат.*

Этот результат с первого взгляда кажется парадоксальным, ибо можно было ожидать, что простое вращение системы координат не может изменять взаимной ориентации базисных векторов, определяющих матрицу. Однако в этом преобразовании матрица A принимает участие именно в качестве *матрицы*, а не как совокупность векторов. Мы рассматривали матрицу A как совокупность векторов, представленных ее столбцами. Хотя эта точка зрения пригодна для геометрической интерпретации системы линейных уравнений, необходимо себе уяснить, что преобразование матрицы *не* следует закону преобразования вектора. Вектор преобразуется, согласно закону

$$\bar{b} = \tilde{U}b, \quad (2-27.1)$$

а матрица, согласно закону

$$\bar{A} = \tilde{U}AU, \quad (2-27.2)$$

где U есть ортогональная матрица

$$\tilde{U}U = I. \quad (2-27.3)$$

Поэтому столбцы матрицы A не остаются инвариантными ни по длине, ни по ориентации, когда она подвергается ортогональному преобразованию. Если мы представляем себе столбцы матрицы A как базисные векторы, то эти векторы в результате ортогонального преобразования изменяют свою длину и ориентацию. Существует, в частности, такое преобразование координат, которое изменяет ориентацию этих осей желательным образом, делая их взаимно ортогональными. Переводя таким образом произвольную косоугольную систему в ортогональную.

Рассмотрим уравнение

$$Ax = b \quad (2-27.4)$$

в той форме, в которой оно решается по способу наименьших квадратов:

$$\tilde{A}Ax = \tilde{A}b \quad (*). \quad (2-27.5)$$

Матрица $\tilde{A}A$ симметрична. Ее собственные значения все вещественны и положительны (или равны нулю), а главные оси ортогональны. Собственные значения μ_i матрицы $\tilde{A}A$ не находятся в прямой связи с собственными значениями λ_i самой матрицы A . Величины λ_i могут быть любыми комплексными числами, тогда как величины μ_i вещественны.

Целесообразнее обозначить собственные значения матрицы $\tilde{A}A$ через μ_i^2 , чтобы явно отметить, что они не только вещественны, но и положительны. Если A сама является симметрической матрицей, то $\tilde{A}A = A^2$. В этом случае собственные значения матрицы $\tilde{A}A$ будут равны λ_i^2 и следовательно, $\mu_i = \lambda_i$.

Но главные оси матрицы $\tilde{A}A$ образуют вещественную ортогональную систему в n -мерном пространстве, которую можно ввести как новую систему координат. В этой новой системе уравнение (5) запишем так:

$$\tilde{\tilde{A}}\tilde{\tilde{A}}\tilde{x} = \tilde{\tilde{A}}\tilde{b}. \quad (2-27.6)$$

Кроме того, в ней $\tilde{\tilde{A}}\tilde{\tilde{A}}$ становится диагональной матрицей **)

$$\tilde{\tilde{A}}\tilde{\tilde{A}} = \begin{bmatrix} \mu_1^2 & & & \\ & \mu_2^2 & & \\ & & \ddots & \\ & & & \mu_n^2 \end{bmatrix}. \quad (2-27.7)$$

*) См. (25.3).

**) Ср. § 9.

Так как, однако, получение матричного произведения $\tilde{A}\bar{A}$ означает умножение по столбцам матрицы на самое себя, то мы получаем для базисных векторов $\bar{a}_i$ новой линейной системы следующие соотношения:

$$\left. \begin{aligned} \bar{a}_i \cdot \bar{a}_k &= 0 & (i \neq k), \\ \bar{a}_i^2 &= \mu_i^2. \end{aligned} \right\} \quad (2-27.8)$$

Таким образом, ортогональность новых базисных векторов доказана, хотя длины этих векторов отнюдь не равны 1. Собственные значения μ_i^2 матрицы $\tilde{A}\bar{A}$ могут быть интерпретированы как квадраты длин новых базисных векторов $\bar{a}_i$.

Составленное нами ортогональное преобразование всегда существует. Некоторые из μ_i^2 могут оказаться равными, и тогда ортогональная матрица U определяется неоднозначно, так как в определенных направлениях имеют место «сферические» условия, и любые две взаимно перпендикулярные оси длиной 1 могут быть взяты в качестве главных осей. Но это означает лишь, что в этих случаях ортогональное преобразование, переводящее первоначальную косоугольную систему осей в прямоугольную, не единственно. В случае особых систем одно или несколько значений μ_i оказываются равными нулю, но в остальном сохраняют силу общие законы. Однако, как легко видеть, в случае особой матрицы *целый столбец преобразованной матрицы $\bar{A}$ становится нулевым*, ибо сумма квадратов элементов столбца может равняться нулю только тогда, когда каждый его элемент равен нулю.

Интересный пример такого преобразования представляет собой некоторый крайний случай, который поучителен тем, что иллюстрирует поведение весьма косоугольных систем. Рассмотрим матрицу

$$A = \begin{bmatrix} a_1 & a_1 & \dots & a_1 \\ a_2 & a_2 & \dots & a_2 \\ \cdot & \cdot & \dots & \cdot \\ a_n & a_n & \dots & a_n \end{bmatrix}.$$

Здесь все оси первоначальной косоугольной системы сливаются в одну, осуществляя, таким образом, наиболее крайний случай линейной зависимости. Матрица $\tilde{A}\bar{A}$ содержит теперь лишь один элемент $\sum a_i^2$, повторенный n^2 раз. Мы можем нормировать его к 1:

$$a_1^2 + a_2^2 + \dots + a_n^2 = 1.$$

Решение задачи собственных значений матрицы $\tilde{A}\bar{A}$ будет

либо $\lambda = n; \quad x_1 = x_2 = \dots = x_n = \frac{1}{\sqrt{n}},$

либо $\lambda = 0; \quad x_1 + x_2 + \dots + x_n = 0.$

Второе решение распадается на $n - 1$ ортогональных осей, ибо собственное значение $\lambda = 0$ имеет кратность $n - 1$. Ортогональная матрица U , столбцы которой дают собственные векторы матрицы $\tilde{A}$, может быть записана следующим образом:

$$U = \begin{bmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2 \cdot 3}} & \cdots & \frac{1}{\sqrt{(n-1)n}} \\ \frac{1}{\sqrt{n}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2 \cdot 3}} & \cdots & \frac{1}{\sqrt{(n-1)n}} \\ \cdots & 0 & -\frac{2}{\sqrt{2 \cdot 3}} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{1}{\sqrt{n}} & 0 & 0 & \cdots & \cdots \\ \frac{1}{\sqrt{n}} & 0 & 0 & \cdots & -\frac{n-1}{\sqrt{(n-1)n}} \end{bmatrix},$$

и, составив произведение $\tilde{U}AU$, мы получаем преобразованную матрицу $\bar{A}$ в новой системе координат:

$$\bar{A} = \begin{bmatrix} a_1 + a_2 + \cdots + a_n, & 0, \dots, 0 \\ \sqrt{n} \frac{a_1 - a_2}{\sqrt{2}}, & 0, \dots, 0 \\ \sqrt{n} \frac{a_1 + a_2 - 2a_3}{\sqrt{2 \cdot 3}}, & 0, \dots, 0 \\ \cdots & \cdots \\ \sqrt{n} \frac{a_1 + a_2 + \cdots - (n-1)a_n}{\sqrt{(n-1)n}}, & 0, \dots, 0 \end{bmatrix}.$$

Условия ортогональности здесь выполняются тривиально, ибо последние $n - 1$ векторов тождественно равны нулю. Но наш пример показывает, что происходит, когда первоначальная матрица A имеет не столь исключительную форму, и не все ее столбцы совпадают, хотя они отличаются друг от друга лишь на небольшие величины. В этом случае место нулей в вышеприведенной матрице $\bar{A}$ занимают элементы малой величины, однако, столбцы остаются ортогональными друг другу. Суммы квадратов элементов дают последовательно $\mu_1^2, \mu_2^2, \dots, \mu_n^2$. В то время как μ_1 опять может быть нормировано к 1 умножением A на общий масштабный множитель, последующие $\mu_2, \mu_3, \dots, \mu_n$ будут малыми числами, если столбцы матрицы все без исключения наклонены под малыми углами друг к другу. Однако столь же возможно, что лишь немного «плохих» осей образуют очень малые углы с подпространством, определяемым остальными осями. Произведение

$$\Delta = \mu_1 \mu_2 \cdots \mu_n$$

равно абсолютной величине определителя матрицы A . Если n велико, а определитель матрицы A очень мал, то это может произойти в двух случаях. Либо относительно большое число множителей μ_i умеренно мало, либо имеется большое число множителей μ_i порядка 1 и несколько чрезвычайно малых чисел μ_i . При любых обстоятельствах новая система координат исключительно удобна как для численной оценки, так и для теоретического изучения обширной системы уравнений.

28. Влияние помех на решение обширных линейных систем

Обращение матриц высокого порядка является кропотливым процессом. Если, соблюдая надлежащую осторожность, нам удалось получить математически удовлетворительное решение, то все же остается открытым вопрос, до какой степени это решение имеет значение для данной физической задачи. Весьма точные вычисления требуют весьма точных данных. Но данные физических задач часто весьма далеки от той точности, которой требуют математические выкладки. В частности, правые части линейных систем являются часто результатом физических измерений, и их точность не может быть гарантирована больше чем до двух-трех значащих цифр. Поэтому обязательно надо исследовать, какое влияние на решение имеют малые, но беспорядочные изменения элементов правой части системы. Это исследование не является простым, если проводить его в первоначальной косоугольной системе координат, в которой заданы последовательные столбцы матрицы A , но мы получим идеально подходящие условия, если введем ту повернутую систему координат, в которой матрица $\tilde{A}A$ имеет диагональную форму.

Преобразование характеризуется уравнениями

$$x = U\bar{x}, \quad \bar{x} = \tilde{U}x. \quad (2-28.1)$$

В новой системе координат матрица A становится ортогональной со столбцами a_i , длины которых суть $\mu_1, \mu_2, \dots, \mu_n$. Решение системы $\tilde{A}\tilde{A}x = \tilde{A}b$ может быть теперь получено следующим образом. В левой части неизвестные разделены

$$\mu_1^2 \bar{x}_1, \mu_2^2 \bar{x}_2, \dots, \mu_n^2 \bar{x}_n. \quad (2-28.2)$$

В правой части нам надо выполнить операцию

$$a_i b = |a_i| |b| \cos \theta_i = \mu_i |b| \cos \theta_i, \quad (2-28.3)$$

где θ_i обозначает угол между векторами b и a . Следовательно,

$$\mu_i^2 \bar{x}_i = \mu_i |b| \cos \theta_i, \quad (2-28.4)$$

что дает

$$\bar{x}_i = \frac{|b| \cos \theta_i}{\mu_i}. \quad (2-28.5)$$

Трудность решения почти особых систем вызывается делением на μ_i , если μ_i очень мало. В результате этого деления решение оказывается весьма чувствительным к небольшим ошибкам в определении вектора b . Истинное положение вектора b неизвестно. Оно маскируется добавочным вектором погрешности, который мал по сравнению с длиной вектора b (рис. 20). Этот вектор погрешности, однако, носит случайный характер и имеет составляющие как в направлении больших, так и в направлении *малых* векторов a_i . Если теперь наименьший вектор a_n

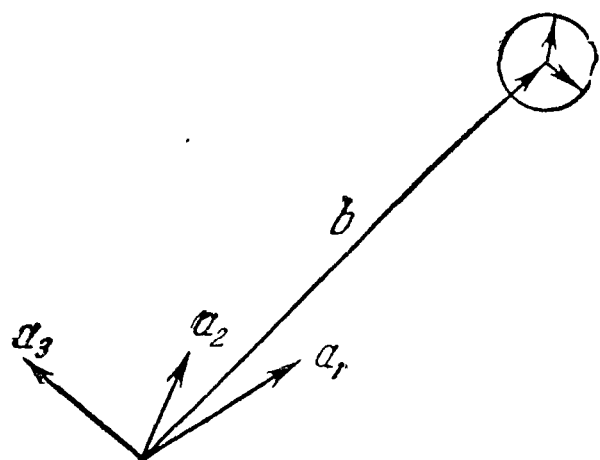


Рис. 20.

имеет длину 10^{-3} сравнительно с длиной наибольшего вектора a_i , т. е. если отношение наименьшего собственного значения матрицы $\tilde{A}A$ к наибольшему составляет 10^{-6} , то при вычислении x_n помеха будет увеличена в 1000 раз. Это легко может привести к такому искажению решения, которое делает его с физической точки зрения негодным. Решение $x_1, x_2, \dots, x_n$ линейной системы часто не благоприятствует весьма ма-

лым собственным векторам матрицы $\tilde{A}A$, но благоприятствует всем собственным векторам одного и того же порядка величины. Это значит, что все $\bar{x}_i$ получаются одного и того же порядка величины. Но тогда вектор b весьма не благоприятствует малым векторам a_i ввиду умножения на малые множители μ_i . Отношение (5) могло бы получиться величиной среднего порядка в связи с тем, что и числитель и знаменатель одновременно становятся малыми. Однако в действительности это не так, ибо в то время как сам вектор b сильно влияет неблагоприятно относительно малых собственных векторов, помеха не сказывается так неблагоприятно в этих направлениях, и поэтому получаются ошибочные составляющие x_i , которые значительно превышают правильные составляющие. В результате может получиться решение с ошибкой на 100% или больше.

Этот анализ показывает, что критической величиной, которая решает вопрос о физической надежности строго математического решения, является не определитель системы, а *отношение наибольшего собственного значения симметрической матрицы $\tilde{A}A$* ¹⁾ к *наименьшему*. Квадратный корень этого отношения измеряет увеличение помех в направлении, соответствующем наименьшему собственному значению. До тех пор пока это отношение не превышает определенную опасную величину, проблема помех не является критической. Но если отношение достигает величины 10^4 и более, увеличение помех в направлении, соответствующем наименьшему собственному значению, становится равным 100 и больше. Точность наших физических измерений

¹⁾ Если A имеет комплексные элементы, то симметрическая матрица $\tilde{A}A$ заменяется эрмитовой матрицей $\tilde{A}^*A$.

редко достаточно, чтобы такое увеличение помех в некоторых направлениях не исказило результаты. Линейная система, при которой критическое отношение превышает 10^4 , едва ли может считаться пригодной для определения искомым неизвестных.

Система координат, составленная из главных осей матрицы $\tilde{A}A$, имеет еще и другое преимущество. Она выявляет в отчетливой форме те отдельные комбинации переменных, которые слишком слабо представлены в заданной системе. В этой системе координат переменные $\bar{x}_i$ полностью уединены (разделены). Кроме того, некоторые $\bar{x}_i$ входят в уравнения со слишком малым множителем. Эти $\bar{x}_i$ могут быть сразу выделены, как те величины, которые практически выпадают из системы уравнений. Но мы можем вернуться к нашим исходным переменным x_i , помня, что переменные, с которыми мы оперировали, находятся в следующем отношении к первоначальным:

$$\bar{x} = \tilde{U}x.$$

Так как матрица U задана, то мы можем найти те линейные комбинации неизвестных, которых не хватает в заданной линейной системе.

В нашей первой очень упрощенной задаче (24.1) мы рассматривали матрицу второго порядка

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1 + \varepsilon \end{bmatrix},$$

где ε — малая величина. Нетрудно решить соответствующую задачу собственных значений для матрицы $\tilde{A}A$. Окончательное решение будет, если считать ε малой величиной:

$$U = \begin{array}{cc} \mu = 2 & \varepsilon/2 \\ \hline \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \end{array}.$$

Составляем уравнения $\bar{x} = \tilde{U}x$:

$$\bar{x}_1 = \frac{1}{\sqrt{2}}(x_1 + x_2), \quad \bar{x}_2 = \frac{1}{\sqrt{2}}(-x_1 + x_2).$$

Малому собственному значению соответствует $\bar{x}_2$. Следовательно, именно линейная комбинация

$$-x_1 + x_2$$

слишком слабо представлена в нашей системе. В этом простом примере мы могли этот факт установить простым обзором заданных уравнений. Но в случае более сложных систем у нас нет простых путей

установления того частного линейного соотношения (или тех соотношений) переменных, которое практически отсутствует в заданной системе. Построение матрицы U дает систематический ответ на этот вопрос, а знание собственных значений μ_1^2 дает количественную меру слабости, с которой эти комбинации входят в заданную систему.

Если при решении задачи мы не входим в подробный анализ влияния помех путем нахождения собственных значений и собственных векторов матрицы $\tilde{A}A$, то все же мы обязательно должны при этом убедиться, что физические помехи не заглушат предполагаемое решение. С этой целью мы изменяем заданную правую часть системы на случайные величины того же порядка, как и ошибки измерений, и определяем влияние этих изменений на полученное решение. Если решение изменяется в результате этого возмущения слишком значительно, то мы должны прийти к заключению, что наше решение, хотя и правильное математически, не может считаться достаточно надежным решением физической задачи.

ЛИТЕРАТУРА К ГЛАВЕ II

- [1] Aitken A. C., Determinants and Matrices (Interscience Publishers, New York, 1944).
- [2] Ferrar W. L., Algebra (Oxford Press, New York, 1941).
- [3] Р. Фрезер, В. Дункан и А. Коллар, Теорит матриц, ИЛ, М., 1950.
- [4] Turnbull H. W., The Theory of Determinants, Matrices and Invariants (Blackie & Son, Glasgow, 1929).
- [5] Каган В. Ф., Основания теории определителей, Гос. изд-во Украины, 1922.
- [6] Гантмахер Ф. Р., Теория матриц, Гостехиздат, 1953.
- [7] Фаддеев Д. К. и Фаддеева Д. Н., Вычислительные методы линейной алгебры, Физматгиз, 1960.
- [8] Дубнов Я. С., Основы векторного исчисления, ч. 2, Гостехиздат, М., 1952.
- [9] Bellman R., Introduction to matrix Analysis, N. Y., 1960.

Статьи

- [10] Лопшиц А. М., Численный метод нахождения собственных значений и собственных плоскостей линейного оператора, труды сем. по вект. и тензорному исчислению, т. VII, 239—259.
-

ГЛАВА III

СИСТЕМЫ МНОГИХ ЛИНЕЙНЫХ УРАВНЕНИЙ

1. Историческое введение

Основатели исчисления бесконечно малых отважились выступить в новом направлении, которое сначала казалось радикально отличным от более консервативных понятий алгебры. Оперирование с «бесконечно малыми» величинами и величинами, «исчезающими в первой, второй, ..., n -й степени», хотя и представляло очевидные преимущества, приводило к серьезным логическим трудностям, которые отметил Г. Беркли (G. Berkeley) в своем *Analyste* (1734). Прежде чем в начале XIX столетия Коши и Гаусс установили точное понятие предела, Лагранж пытался найти другое решение. Он показал, как можно, применяя понятия алгебры, решать также и задачи высшей математики. Мы не можем отличить производные от обычных разностных отношений, если Δx взято достаточно малым, а интегрирование может быть заменено суммированием.

Эта тенденция «алгебраизировать» исчисление бесконечно малых и все процессы высшей математики приобретала все возрастающее значение и, наконец, революционизировала все здание высшей математики. Фредгольм (1900) дал строгое обоснование всей теории, показав, что один класс интегральных уравнений может быть представлен как предел обыкновенной системы линейных алгебраических уравнений, порядок которой возрастает до бесконечности. Более того, ошибка становится малой уже тогда, когда порядок соответствующей алгебраической системы еще не очень велик.

Это направление в математике носило вначале чисто теоретический характер; в наши дни оно приобрело большое практическое значение в связи с созданием электронных вычислительных машин, которые сделали возможным решение обширных систем алгебраических уравнений. Сведение задачи с граничными условиями к решению обширной системы алгебраических уравнений уже не является поэтому только методом чисто теоретического исследования, оно стало весьма эффективным практическим средством решения дифференциальных уравнений в частных производных.

Однако фактическое обращение «большой матрицы» даже теперь является устрашающе трудным делом. Вместо получения точного решения с помощью обращения матрицы предпочтительно применять более простые приемы, которые отдельными этапами приводят все ближе и ближе к искомому решению. Из таких этапов образуется «итерационная техника», основанная на постоянном повторении одного и того же алгоритма. Такого рода операции особенно удобно программировать на вычислительных машинах. Простейшее и наиболее естественное итерирование использует последовательные степени матрицы. Мы, таким образом, приходим к необходимости исследования полиномов, образованных с помощью матриц.

2. Операции с матричными полиномами

Если x есть какая-либо алгебраическая величина, то повторные операции умножения и сложения приводят к линейной комбинации степеней x , называемой «полиномом от x »:

$$P_n(x) = p_n x^n + p_{n-1} x^{n-1} + \dots + p_1 x + p_0. \quad (3-2.1)$$

Полиномы от x имеют исключительно многообразное применение благодаря их гибкости и той легкости, с которой над ними можно производить такие аналитические операции, как дифференцирование и интегрирование. Считая x некоторой матрицей A , мы исследуем возможность применения матричных полиномов для решения некоторых матричных задач. Матричная алгебра является глубоким аналогом обыкновенной алгебры, хотя в ней и приходится жертвовать коммутативным законом умножения. Но если оперировать с одной только матрицей, в предположении, что коэффициенты полинома $P_n(A)$ суть обыкновенные числа, то некоммутативный характер матриц не играет роли, ибо каждая матрица коммутирует сама с собой*). Следовательно, имеет смысл исследовать возможные преимущества матричных полиномов

$$P_n(A) = p_n A^n + p_{n-1} A^{n-1} + \dots + p_1 A + p_0. \quad (3-2.2)$$

Однако с точки зрения приложений мы тотчас же встречаемся со следующим затруднением. Последовательные степени x образуются с помощью простого умножения предыдущей степени x^{k-1} на x . Подобным же образом можно получать последовательные степени матрицы:

$$A^k = A A^{k-1}.$$

Но эта операция требует выполнения потрясающего количества отдельных действий над числами, ибо умножение на матрицу осуществляется почленным перемножением строк на столбцы сомножителей. Это при-

*) A также и со своей произвольной степенью.

Природа операции $P_n(A)b_0$ лучше всего может быть прослежена, если использовать собственные векторы матрицы A (ср. §§ 6 и 7):

$$Au = \lambda u, \quad \tilde{A}v = \lambda v. \quad (3-2.10)$$

Первое из этих уравнений имеет n решений (если считать, что A есть «полноосная» матрица):

$$u_1, u_2, \dots, u_n.$$

Второе уравнение имеет n решений:

$$v_1, v_2, \dots, v_n.$$

Эти две системы векторов находятся в биортогональном соотношении между собой

$$\left. \begin{aligned} u_i v_k &= 0 & (\text{при } i \neq k), \\ u_i v_i &= 1. \end{aligned} \right\} \quad (3-2.11)$$

Разложим вектор b_0 по осям базисных векторов

$$b_0 = \beta_1 u_1 + \beta_2 u_2 + \dots + \beta_n u_n, \quad (3-2.12)$$

где $\beta_i = b_0 v_i$. Тогда в силу определения главных осей, мы получаем

$$\begin{aligned} Ab_0 &= \beta_1 \lambda_1 u_1 + \beta_2 \lambda_2 u_2 + \dots + \beta_n \lambda_n u_n, \\ A^k b_0 &= \beta_1 \lambda_1^k u_1 + \beta_2 \lambda_2^k u_2 + \dots + \beta_n \lambda_n^k u_n, \end{aligned}$$

а для произвольного полиномиального оператора

$$P_n(A)b_0 = \beta_1 P_n(\lambda_1)u_1 + \beta_2 P_n(\lambda_2)u_2 + \dots + \beta_n P_n(\lambda_n)u_n. \quad (3-2.13)$$

Мы видим, таким образом, что после введения собственных векторов матрицы A в качестве новой системы координат действие матричных полиномов $P_n(A)$ сводится к действию чисто алгебраических полиномов $P_n(\lambda)$, где λ должно замещаться последовательными собственными значениями $\lambda_1, \lambda_2, \dots, \lambda_n$ матрицы A .

3. p, q -алгоритм

Особенно полезная система полиномов образуется самой матрицей A , если мы вместе с ней рассмотрим произвольный вектор b , играющий роль «правой части» линейного уравнения $Ax = b$, порожденного заданной матрицей. Для итерационного метода, который мы рассмотрим ниже, вектор b отождествляется с «исходным вектором» b_0 (координаты b_0 суть произвольно выбранные числа), на который должна действовать матрица A . Мы можем составить последовательные «ите-рации» $A^k b_0 = b_k$, образуя последовательно

$$b_0, b_1 = Ab_0, b_2 = Ab_1 = A^2 b_0, \dots, b_n = Ab_{n-1} = A^n b_0. \quad (3-3.1)$$

В силу тождества Гамильтона — Кели последний из этих векторов является линейной комбинацией предыдущих. Хотя, вообще говоря, предыдущие векторы линейно независимы, мы можем на каждом этапе этого процесса искать «наилучшее тождество», которое можно установить между ними. Это значит, что хотя линейная комбинация

$$p_k = b_k + \gamma_{k-1}b_{k-1} + \dots + \gamma_0b_0 \quad (3-3.2)$$

не может быть сделана равной нулю ни при каком выборе коэффициентов γ_k^*), мы можем тем не менее путем соответствующего выбора чисел γ_i минимизировать длину этого вектора. Если A — симметрическая матрица, то мы ищем минимум скаляров

$$p_k^2 = (b_k + \gamma_{k-1}b_{k-1} + \dots + \gamma_0b_0)^2. \quad (3-3.3)$$

Если A не симметрична, то следует отличать действие самой матрицы A от действия матрицы, сопряженной к ней. Целесообразно поэтому дополнить систему векторов (3.1) сопряженной ей системой векторов

$$\tilde{b}_0, \tilde{b}_1 = \tilde{A}b_0, \dots, \tilde{b}_n = \tilde{A}b_{n-1}. \quad (3-3.4)$$

и искать минимум величины, определяемой скаляром

$$p_k \cdot \tilde{p}_k = (b_k + \gamma_{k-1}b_{k-1} + \dots + \gamma_0b_0)(\tilde{b}_k + \gamma_{k-1}\tilde{b}_{k-1} + \dots + \gamma_0\tilde{b}_0). \quad (3-3.5)$$

Решение этой экстремальной задачи приводит к системе линейных уравнений для коэффициентов γ_k , характеризующейся особым рода матрицей, которую мы назовем «рекуррентной». Однако вместо решения этой системы независимо для каждого k мы можем развить легко программируемый «прогрессивный алгоритм»¹⁾ для постепенного образования полиномов

$$p_k(A)b_0 = (A^k + \gamma_{k-1}A^{k-1} + \dots + \gamma_0)b_0 \quad (3-3.6)$$

и второй связанной с ними системы полиномов $q_k(A)b_0$. Когда мы достигнем, наконец, значения $k=n$, то получим полином $p_n(A)b_0$, тождественно равный нулю. Полином $p_n(A)$ совпадает с характери-

*) Если, конечно, векторы $b_0, b_1, \dots, b_k$ линейно независимы. Однако векторы $b_0, Ab_0, \dots, A^kb_0$ могут оказаться и линейно зависимыми. Это случится, например, уже для $k=1$, если b_0 есть собственный вектор матрицы A , для $k=2$ — если b_0 есть линейная комбинация двух различных собственных векторов и т. д.

Линейная зависимость векторов $b_1, \dots, b_k$ может иметь место и для любого исходного вектора b_0 в том случае, когда матрица A имеет специальную структуру. Например, если

$$A = \alpha I, \quad \text{то} \quad b_1 = \alpha b_0.$$

¹⁾ Ср. статью [5] литературного указателя, которая содержит исчерпывающий разбор многочисленных интересных свойств этих полиномов.

стическим уравнением матрицы, его корни дают все собственные значения для A .

Полиномы $p_k(A)$ (а также $q_k(A)$) относятся к типу классических ортогональных полиномов. Они удовлетворяют рекуррентному соотношению вида (ср. V, 21)

$$p_{k+1}(A) = (A - \alpha_k) p_k(A) - \beta_k p_{k-1}(A). \quad (3-3.7)$$

Числовые константы α_k и β_k этих соотношений не могут быть заранее заданы; они определяются самой матрицей в соединении с правой частью b_0 и выявляются постепенно по мере того, как разворачивается алгоритм. Этот, так называемый « p, q -алгоритм» дает полное решение задачи нахождения собственных значений. Собственные значения и собственные векторы возникают в нем путем последовательных приближений, которые все время улучшаются и в конце процесса становятся точными решениями (если не считать ошибок округления). Особенно важны начальные этапы процесса. Эта фаза алгоритма заслуживает внимания и сама по себе, так как имеет полезные приложения. Вычислим три пары векторов:

$$\begin{array}{ccc} b_0, & b_1, & b_2, \\ \tilde{b}_0, & \tilde{b}_1, & \tilde{b}_2 \end{array}$$

и образуем скалярные произведения *)

$$\left. \begin{array}{l} c_0 = b_0 \tilde{b}_0, \\ c_1 = b_0 \tilde{b}_1 = b_1 \tilde{b}_0, \\ c_2 = b_0 \tilde{b}_2 = b_1 \tilde{b}_1 = b_2 \tilde{b}_0, \\ c_3 = b_1 \tilde{b}_2 = b_2 \tilde{b}_1. \end{array} \right\} \quad (3-3.8)$$

*) Числа c_0, c_1, c_2, c_3 рассматриваются потому, что они сразу возникают при решении задачи о нахождении таких коэффициентов γ_0, γ_1 , при которых делается минимальным скалярное произведение (5) в случае, когда $k=2$; легко убедиться в том, что квадратичное выражение

$$\begin{aligned} p_2 \tilde{p}_2 &= (b_2 + \gamma_1 b_1 + \gamma_0 b_0) \cdot (\tilde{b}_2 + \gamma_1 \tilde{b}_1 + \gamma_0 \tilde{b}_0) = \\ &= \gamma_0^2 c_0 + 2\gamma_0 \gamma_1 c_1 + 2\gamma_0 c_2 + 2\gamma_1 c_3 + c_4, \quad [c_4 = b_0 A^2 b_0] \end{aligned}$$

принимает минимальное значение при следующих значениях γ_0 и γ_1 :

$$\gamma_0 = \frac{\begin{vmatrix} c_1 & c_2 \\ c_2 & c_3 \end{vmatrix}}{\begin{vmatrix} c_0 & c_1 \\ c_1 & c_2 \end{vmatrix}}; \quad \gamma_1 = \frac{\begin{vmatrix} c_2 & c_0 \\ c_3 & c_1 \end{vmatrix}}{\begin{vmatrix} c_0 & c_1 \\ c_1 & c_2 \end{vmatrix}}.$$

Отсюда следует, что вектор

$$\left(\begin{vmatrix} c_0 & c_1 \\ c_1 & c_2 \end{vmatrix} A^2 + \begin{vmatrix} c_2 & c_0 \\ c_3 & c_1 \end{vmatrix} A + \begin{vmatrix} c_1 & c_2 \\ c_2 & c_3 \end{vmatrix} \right) b_0$$

хотя и не равен, вообще говоря, нулю, но отличается от нуля «в минималь-

(Если $b_0 = \tilde{b}_0$, то в случае симметрической матрицы $\tilde{b}_k = b_k$, в то время как в случае эрмитовых матриц $\tilde{b}_k = b_k^*$.) Можно доказать, что уравнение*)

$$\begin{vmatrix} 1 & \lambda & \lambda^2 \\ c_0 & c_1 & c_2 \\ c_1 & c_2 & c_3 \end{vmatrix} = 0 \quad (3-3.9)$$

дает два корня, близкие к двум наибольшим собственным значениям матрицы A , если только мы возьмем такой исходный вектор b_0 , направление которого благоприятствует получению хорошего приближения. Такой вектор можно получить, умножив произвольные случайно выбранные векторы $v_0, \tilde{v}_0$ на $A, \tilde{A}$ и их последовательные степени, т. е. образуя последовательно векторы

$$\begin{aligned} v_1 &= Av_0, & v_2 &= Av_1, & \dots, & v_m &= Av_{m-1}, \\ \tilde{v}_1 &= \tilde{A}\tilde{v}_0, & \tilde{v}_2 &= \tilde{A}\tilde{v}_1, & \dots, & \tilde{v}_m &= \tilde{A}\tilde{v}_{m-1} \end{aligned}$$

и положив затем $b_0 = v_m, \tilde{b}_0 = \tilde{v}_m$. Если m равно примерно 5 или 6, то наибольшие собственные значения будут уже играть основную роль**). Решив уравнение (9), мы получим теперь с хорошим приближением наибольшее собственное число симметрической матрицы или наибольшую по модулю пару сопряженных комплексных собственных чисел произвольной матрицы.

Этот метод представляет собой естественное распространение метода Бернулли для нахождения наибольших корней алгебраического уравнения [ср. (1-15.10)]. Этот же метод применим к задаче (часто даже более важной) нахождения наименьших собственных значений матрицы. Следует только воспользоваться преобразованием, которое обращает порядок модулей собственных значений λ_i (см. §§ 9 и 10). Собственные векторы, соответствующие двум наибольшим по модулю корням, также теперь получатся с хорошим приближением.

ной мере», и поэтому имеет место следующее приближенное равенство:

$$(A - \lambda_1 I)(A - \lambda_2 I)b_0 = 0, \quad (3-3.9')$$

в котором λ_1 и λ_2 — корни уравнения (9').

*) Легко убедиться, что его левая часть представляет собой скалярное произведение $(\tilde{A} - \lambda)b_0(A - \lambda)\tilde{b}_0$.

**) Действительно, в силу (2.13)

$$A^m v_0 = \beta_1 \lambda_1^m u_1 + \beta_2 \lambda_2^m u_2 + \dots + \beta_n \lambda_n^m u_n,$$

и, следовательно, если модули чисел λ_1 и λ_2 больше, чем модули остальных собственных чисел $\lambda_3, \lambda_4, \dots, \lambda_n$, то первые два слагаемых дают основную часть значения вектора $A^m v_0$.

Так как уравнение (9) означает, что приближенно выполняется соотношение *)

$$(A - \lambda_1 I)(A - \lambda_2 I)b_0 = 0, \quad (3-3.9')$$

то мы получаем для собственного вектора u_1 (соответствующего собственному числу λ_1) следующее выражение **):

$$u_1 = Ab_0 - \lambda_2 b_0 = b_1 - \lambda_2 b_0$$

или, в лучшем приближении, умножая на A ,

$$u_1 = b_2 - \lambda_2 b_1 \quad (3-3.10)$$

и аналогично для собственного вектора u_2 (соответствующего собственному числу λ_2):

$$u_2 = b_2 - \lambda_1 b_1. \quad (3-3.11)$$

Подобным же образом для сопряженных векторов получаем

$$\tilde{u}_1 = \tilde{b}_2 - \lambda_2 \tilde{b}_1, \quad \tilde{u}_2 = \tilde{b}_2 - \lambda_1 \tilde{b}_1. \quad (3-3.12)$$

(Точность, с которой получаются векторы u_2 , $\tilde{u}_2$, значительно меньше точности векторов u_1 , $\tilde{u}_1$ если абсолютное значение λ_1 значительно превышает значение λ_2 ; однако в случае сопряженных комплексных собственных значений оба вектора имеют одинаковую степень точности, так как $|\lambda_1| = |\lambda_2|$.) ***)

4. Полиномы Чебышева

В предыдущем параграфе мы занимались полиномами, порожденными самой матрицей A . Коэффициенты α_k , β_k рекуррентной формулы (3.7) надо было получать последовательными этапами в процессе минимизирующего процесса. Значительно проще те процессы, которые используют универсальные полиномы, для которых коэффициенты α_k , β_k представляют собой некоторые наперед заданные постоянные. Среди этих полиномов простейшими и наиболее полезными являются полиномы Чебышева, названные так по имени русского математика П. Л. Чебышева (1821—1894). Они имеют то преимущество, что для них α_k , β_k суть постоянные, не зависящие от k . Поэтому таблицы этих полиномов составляются особенно просто.

*) См. примечание на стр. 188.

**) Ибо (9') можно записать так:

$$(A - \lambda_1 I)(Ab_0 - \lambda_2 b_0) = 0.$$

***) Другие способы использования последовательных степеней матрицы для нахождения ее собственных векторов и собственных чисел, пригодные и в том случае, когда матрица имеет сложную структуру, изложены в статьях [7] и [8].

В этой книге мы встретим эти замечательные полиномы в самых различных формах (ср. IV, 16; V, 20; гл. VII). Сейчас нас интересует их применение к операциям с матрицами. Эти полиномы обладают большим удобством благодаря той особенно простой рекуррентной формуле, с помощью которой они могут быть образованы.

Мы исходим из простого тригонометрического тождества

$$\cos(n+1)\theta + \cos(n-1)\theta = 2\cos\theta\cos n\theta \quad (3-4.1)$$

и аналогичного ему

$$\sin(n+1)\theta + \sin(n-1)\theta = 2\cos\theta\sin n\theta. \quad (3-4.2)$$

Эти тождества дают возможность, зная значения $\cos\theta$ и $\sin\theta$, вычислить все последующие значения $\cos n\theta$ и $\sin n\theta$ последовательными этапами. Действительно, получив значение $\cos n\theta$, мы из тождества (1) можем непосредственно найти $\cos(n+1)\theta$. Это же верно по отношению к $\sin n\theta$ и $\sin(n+1)\theta$ ¹⁾.

Если мы теперь положим $\cos\theta = x$ и начнем с $\cos 0 = 1$ и $\cos\theta = x$, то рекуррентная формула (1) даст для $\cos 2\theta$ квадратный многочлен от x , затем для $\cos 3\theta$ — кубический многочлен и так далее. Мы можем поэтому переписать формулу (1) в следующей форме:

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), \quad (3-4.3)$$

определив тем самым систему полиномов. По этой формуле мы получаем всё те же значения $\cos n\theta$, но они выражены в виде полиномов n -й степени относительно переменной $x = \cos\theta$:

$$\cos n\theta = T_n(x). \quad (3-4.4)$$

Совершенно аналогичное положение возникает для рекуррентной формулы (2), если разделить обе ее части на $\sin\theta$. Теперь мы исходим из тождеств

$$\frac{\sin\theta}{\sin\theta} = 1, \quad \frac{\sin 2\theta}{\sin\theta} = 2\cos\theta$$

и получаем $\frac{\sin 3\theta}{\sin\theta}$ в виде квадратного многочлена относительно x , затем $\frac{\sin 4\theta}{\sin\theta}$ в виде кубического многочлена относительно x и т. д.

Вообще мы получаем теперь значение $\frac{\sin(n+1)\theta}{\sin\theta}$, представленное в виде полинома n -й степени относительно переменной $x = \cos\theta$:

$$\frac{\sin(n+1)\theta}{\sin\theta} = U_n(x). \quad (3-4.5)$$

¹⁾ Интересно отметить, что в первых тригонометрических таблицах, построенных индусами, значения $\sin x$ действительно вычислялись через каждый градус с помощью этой рекуррентной формулы, начиная с $\sin 1^\circ$, как ключевого значения, которое было получено из других соображений.

Рекуррентная формула для $U_n(x)$:

$$U_{n+1}(x) = 2xU_n(x) - U_{n-1}(x) \quad (3-4.6)$$

имеет тот же вид, что и (3), только исходные данные другие. Теперь

$$U_0(x) = 1, \quad U_1(x) = 2x \quad (3-4.7)$$

вместо

$$T_0(x) = 0, \quad T_1(x) = x. \quad (3-4.8)$$

Соотношение

$$x = \cos \theta \quad (3-4.9)$$

имеет простой геометрический смысл. Значения x расположены на отрезке от -1 до $+1$. Строим окружность на этом отрезке, как на диаметре. Тогда θ есть центральный угол, соответствующий той точке P' на окружности, которая проектируется на ось x -ов в точку с абсциссой x (рис. 21).

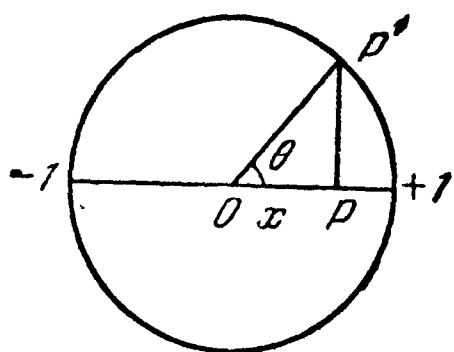


Рис. 21.

В некоторых случаях нас интересует, однако, изменение x лишь в интервале $(0, 1)$. В этом случае мы полагаем

$$x = \frac{1 - \cos \theta}{2} = \sin^2 \frac{\theta}{2} \quad (3-4.10)$$

и определяем «смещенные полиномы Чебышева» $T_n^*(x)$ по рекуррентной формуле

$$T_{n+1}^*(x) = 2(1 - 2x)T_n^*(x) - T_{n-1}^*(x), \quad (3-4.11)$$

исходя из начальных условий

$$T_0^*(x) = 0, \quad T_1^*(x) = 1 - 2x.$$

Аналогично определяются «смещенные полиномы Чебышева» второго рода с помощью той же рекуррентной формулы

$$U_{n+1}^*(x) = 2(1 - 2x)U_n^*(x) - U_{n-1}^*(x), \quad (3-4.12)$$

но исходя из начальных условий

$$U_0^*(x) = 1, \quad U_1^*(x) = 2(1 - 2x). \quad (3-4.13)$$

5. Спектроскопический анализ собственных значений

В проблемах теории вибрации характеристические частоты упругой системы в большинстве случаев определяются собственными значениями заданного дифференциального оператора. Этот оператор аппроксимируют конечным разностным оператором и тогда проблема сводится к матричной задаче в пространстве конечного числа измерений. Особенный интерес с физической точки зрения представляют малые соб-

ственные значения этого оператора. В области больших собственных значений (большие частоты) ошибка, вызванная заменой дифференциального оператора разностным, становится слишком заметной. Кроме того, колебания, которые характеризуются собственными векторами, соответствующими большим собственным значениям, обычно имеют очень малые амплитуды и поэтому практически менее важны.

В p, q -алгоритме, рассмотренном в § 3, собственные значения возникают постепенно, начиная от больших к меньшим. Можно сравнительно быстро получить таким способом большие собственные значения матрицы, однако малые собственные значения получаются лишь к концу тщательно проведенного процесса ортогонализации. А между тем часто нас интересуют преимущественно малые собственные значения и соответствующие собственные векторы.

Одним из возможных путей преодоления этого трудного положения является предварительное обращение матрицы; однако в случае матрицы высокого порядка — это нелегкая работа. Новую перспективу в анализе собственных значений симметрических матриц или общих произвольных матриц, собственные значения которых вещественны, открывают полиномы Чебышева. Мы получаем возможность разыскивать собственные значения матрицы в каком угодно их интервале подобно тому, как спектроскоп может исследовать линейный спектр колеблющегося атома в любом выделенном интервале частот. Мы разлагаем «спектр собственных значений» в спектральные линии, которые можно сделать практически независимыми друг от друга. Обычное «поглощение» малых собственных значений большими устраняется. Это достигается тем, что собственные значения «проектируются на окружность». На этой окружности понятие «большой» и «малый» теряет свое значение. Любая точка окружности эквивалентна другой, ибо всякая точка может быть принята в качестве начала отсчета на окружности.

Из формулы

$$x = \cos \theta \quad (3-5.1)$$

следует, что полиномы Чебышева заданы в интервале $[-1, +1]$ значений x . Чтобы обеспечить в интересующей нас задаче этот интервал, необходимо надлежащим образом нормировать собственные значения матрицы A . Такую нормировку можно произвести с помощью теоремы Гершгорна [17], которая устанавливает верхнюю границу наибольшего по абсолютной величине собственного значения $\lambda = \lambda_M$ матрицы. Теорема эта основана на следующих соображениях. Мы исходим из системы уравнений, определяющих произвольное собственное значение λ и соответствующий ему собственный вектор:

$$a_{i1}x_1 + \dots + a_{in}x_n = \lambda x_i. \quad (3-5.2)$$

То решение, которое соответствует λ_M , мы обозначим через

$$u_M = (\xi_1, \xi_2, \dots, \xi_n).$$

Среди этих (вообще говоря, комплексных) координат ξ_i мы выбираем наибольшую по абсолютной величине ξ_α . Тогда то из уравнений системы (2), которое соответствует $i = \alpha$, дает

$$\lambda_M = a_{\alpha 1} \frac{\xi_1}{\xi_\alpha} + a_{\alpha 2} \frac{\xi_2}{\xi_\alpha} + \dots + a_{\alpha n} \frac{\xi_n}{\xi_\alpha}. \quad (3-5.3)$$

Как известно, абсолютное значение суммы комплексных чисел не может превысить суммы абсолютных значений этих чисел. Поэтому

$$|\lambda_M| \leq |a_{\alpha 1}| \left| \frac{\xi_1}{\xi_\alpha} \right| + |a_{\alpha 2}| \left| \frac{\xi_2}{\xi_\alpha} \right| + \dots + |a_{\alpha n}| \left| \frac{\xi_n}{\xi_\alpha} \right|. \quad (3-5.4)$$

Но вторые множители здесь — числа, лежащие между 0 и 1, и поэтому

$$|\lambda_M| \leq |a_{\alpha 1}| + |a_{\alpha 2}| + \dots + |a_{\alpha n}|. \quad (3-5.5)$$

Составим теперь сумму абсолютных значений элементов каждой строки:

$$|a_{i1}| + |a_{i2}| + \dots + |a_{in}| = s_i. \quad (3-5.6)$$

Мы получим n положительных чисел s_i . Обозначим наибольшее из них через s . Тогда $s_i \leq s$ для всех i , и поэтому, согласно неравенству (5),

$$|\lambda_M| \leq s_\alpha \leq s. \quad (3-5.7)$$

Мы нашли, таким образом, определенную верхнюю границу для наибольшего по абсолютной величине собственного значения произвольной матрицы путем образования сумм абсолютных величин элементов каждой строки и выбора наибольшего из этих положительных чисел. Так как транспонированная матрица имеет те же собственные значения, что и первоначальная, то мы можем ту же операцию произвести над столбцами вместо строк. Наименьший из этих двух максимумов даст лучшую верхнюю границу для $|\lambda_M|$.

Оценка Гершгорна всегда надежна, но часто непрактична, так как мы можем значительно переоценить наибольшее собственное значение. Более реалистическая, но не всегда надежная оценка возможна на основе приема, указанного в § 3 (уравнение (3.9)), дающего приближенное значение наибольшего собственного значения произвольной матрицы.

Допустим теперь, что у нас имеется матрица A , относительно которой известно, что ее собственные числа вещественны, но могут быть положительны или отрицательны. В этом случае мы будем оперировать со следующей «калиброванной» матрицей:

$$C = \frac{2}{s} A. \quad (3-5.8)$$

Собственные числа матрицы C лежат в интервале $(-2, 2)$. С другой стороны, допустим, что все собственные значения матрицы A положительны или равны нулю. В этом случае мы положим

$$C = 2I - \frac{4}{s} A. \quad (3-5.9)$$

Собственные значения C опять будут лежать в интервале $(-2, 2)$. Начинаем с «пробного вектора» b_0 , координаты которого — случайно взятые числа, и образуем последовательность векторов, определяемых рекуррентной формулой [ср. (4.3)]

$$b_{k+1} = Cb_k - b_{k-1} \quad (3-5.10)$$

при исходных значениях

$$b_0 = b_0, \quad b_1 = \frac{1}{2} Cb_0. \quad (3-5.11)$$

Предположим, что пробный вектор b_0 разложен по осям координатной системы, образуемой собственными векторами матрицы A :

$$b_0 = \beta_1 u_1 + \beta_2 u_2 + \dots + \beta_n u_n. \quad (3-5.12)$$

Тогда произвольный вектор b_k из последовательности, образованной по рекуррентным формулам (10), (11), будет равен

$$b_k = \beta_1 (\cos k\theta_1) u_1 + \beta_2 (\cos k\theta_2) u_2 + \dots + \beta_n (\cos k\theta_n) u_n, \quad (3-5.13)$$

где углы $\theta_1, \theta_2, \dots, \theta_n$ соответствуют собственным значениям λ_i матрицы C . В случае (8) это соответствие определяется формулой

$$\lambda_i = \cos \theta_i, \quad (3-5.14)$$

тогда как в случае (9) — формулой

$$\lambda_i = \frac{1 - \cos \theta_i}{2} = \sin^2 \frac{\theta_i}{2}. \quad (3-5.15)$$

Введем теперь следующую векторную функцию $f(t)$ от переменного t :

$$f(t) = (\beta_1 \cos \theta_1 t) u_1 + (\beta_2 \cos \theta_2 t) u_2 + \dots + (\beta_n \cos \theta_n t) u_n. \quad (3-5.16)$$

Легко видеть, что векторы b_k представляют собой значения функции $f(t)$ в целых точках $t = k$:

$$b_k = f(k). \quad (3-5.17)$$

Наша задача может быть теперь сформулирована следующим образом: «Исследуется функция $f(t)$, составленная из чисто периодических слагаемых, имеющих различные амплитуды и различные частоты. Эта функция наблюдается в равностоящие моменты времени

$$t = 0, 1, 2, \dots, N,$$

т. е. заданы только ее значения (17). Найти неизвестные частоты θ_i каждого отдельного периодического слагаемого».

Такого рода задача встречается в гидрологических исследованиях, в метеорологии, астрономии, в экономических исследованиях и в других областях, где задается результирующее взаимодействие некоторых периодических компонент и цель состоит в том, чтобы выделить эти компоненты и найти амплитуду и частоту каждого из составляющих колебаний в отдельности. Эта задача часто называется «разысканием скрытых периодичностей» (ср. IV, 22). Ее решение находится с помощью «преобразования Фурье» (ср. IV, 17). Мы преобразуем первоначальную функцию $f(t)$ в новую функцию $F(p)$ с помощью преобразования Фурье, которое в нашем случае принимает форму суммы, вместо интеграла:

$$F(p) = \frac{1}{2} f(0) + f(1) \cos \frac{\pi}{N} p + f(2) \cos \frac{2\pi}{N} p + \dots \\ \dots + \frac{1}{2} f(N) \cos \frac{N\pi}{N} p. \quad (3-5.18)$$

Эта функция обнаруживает существование периодического компонента $\cos \theta_i t$ тем, что она принимает пиковое значение в точке

$$p = \frac{N}{\pi} \theta_i, \quad (3-5.19)$$

ибо, если $f(t)$ имеет форму $\cos \theta_i t$, то соответствующая функция $F(p)$ будет

$$F(p) = K\left(\frac{N}{\pi} \theta_i + p\right) + K\left(\frac{N}{\pi} \theta_i - p\right), \quad (3-5.20)$$

где

$$K(\xi) = \frac{1}{4} \sin \pi \xi \operatorname{ctg} \frac{\pi \xi}{2N}. \quad (3-5.21)$$

Функция $K(\xi)$ есть функция одного только ξ , и для малых значений ξ мы можем считать, что

$$K(\xi) = \frac{N}{2} \frac{\sin \pi \xi}{\pi \xi}. \quad (3-5.22)$$

Высота пика возрастает равномерно с N , но форма графика функции $F(p)$ в окрестности пика $\xi = 0$ не меняется вместе с N . Так как θ_i лежит между 0 и π , то p лежит между 0 и N ; $F(p)$ есть четная функция от p , и поэтому нет нужды рассматривать отрицательные значения p . Множитель N в правой части равенства (19) показывает, что «разрешающая способность» метода возрастает линейно с N , т. е. с числом итераций. В то время как форма выступа, который характеризует некоторый максимум функции $F(p)$, не изменяется, острота пика пропорциональна N , так как вместе с возрастанием N пики раздвигаются.

гаются и в конце концов даже весьма близкие пики могут быть отделены друг от друга.

Этот метод нахождения собственных значений матрицы путем отыскания скрытых периодичностей суммы периодических функций может быть назван *«спектроскопическим методом»*, так как он математически имитирует действие спектроскопа. Спектроскоп вскрывает частоты, из которых состоит световое излучение возбужденного атома. Эти частоты могут быть вычислены из положения «спектральных линий». Спектральные линии не являются линиями в математическом смысле, но имеют некоторую конечную ширину. Закон распределения амплитуд в окрестности пика аналогичен тому, который дается функцией $F(p)$. Единственное отличие состоит в том, что исключительная точность спектроскопических измерений обуславливается большим постоянством оптических колебаний, которое соответствует весьма большому значению величины N , а мы можем сократить число итераций и все же сохранить высокую точность ввиду большого числа значащих цифр, которыми мы располагаем. Хотя пики теперь расположены значительно теснее, мы можем все же скорректировать наши предварительные результаты, рассчитав интерференцию соседних пиков и вычтя их влияние. Этим путем можно уточнить положение максимумов с относительной точностью в 10^6 и более.

Нет надобности выписывать полностью последовательные векторы $b_0, b_1, \dots, b_m$. Достаточно отметить по одной координате каждого вектора, лишь бы это была постоянно координата с одним и тем же номером (например, первая, вторая, ...) каждого вектора. Мы получаем, таким образом, одномерную последовательность чисел

$$\gamma_0, \gamma_1, \gamma_2, \dots, \gamma_N \quad (3-5.23)$$

которые будут коэффициентами разложения в ряд Фурье по косинусам. Особенно интересны целые значения числа $p=k$. Соответствующие значения функции $F(k)$ обозначим через y_k :

$$y_k = \frac{1}{2} \gamma_0 + \gamma_1 \cos \frac{\pi k}{N} + \gamma_2 \cos \frac{2\pi k}{N} + \dots + \frac{1}{2} \gamma_N \cos \frac{N\pi k}{N}. \quad (3-5.24)$$

Вычисляя эти y_k систематически для $k=0, 1, 2, \dots, N$, преобразуем первоначальную последовательность (23) в новую последовательность

$$y_0, y_1, y_2, \dots, y_N. \quad (3-5.25)$$

Введем обозначение

$$\omega_i = \frac{N}{\pi} \theta_i. \quad (3-5.26)$$

Тогда равенства (20) и (22) дают

$$K(\omega_i - p) = \frac{N}{2} \frac{\sin \pi (\omega_i - p)}{\pi (\omega_i - p)}.$$

Для целых значений p в числителе получается $(-1)^k \sin \pi \omega_i$. Отсюда видно, что интерференция одного пика с другим, если отвлечься от сменяющихся знаков плюс — минус, задается функцией

$$\frac{\sin \pi \omega_i}{\pi (\omega_i - \omega_j)}. \quad (3-5.27)$$

Если случайно некоторый пик при $p = \omega_i$ «попадает» на целое значение $\omega_i = k$, то получается одиночный пик без каких-либо «хвостов». Отдельный максимум с обеих сторон сопровождается нулями. Если, однако, максимум не попадает на целое значение, то максимальная амплитуда сопровождается с обеих сторон меньшими амплитудами. Наименьший спад получается, если максимум находится в точности на полпути между двумя целыми значениями p . Схема амплитуд тогда будет:

$$\frac{1}{9}, -\frac{1}{7}, \frac{1}{5}, -\frac{1}{3}, 1, 1, -\frac{1}{3}, \frac{1}{5}, -\frac{1}{7}, \dots$$

Этот слабый спад амплитуд может быть значительно ускорен, если оперировать со вторыми разностями*) первоначальных амплитуд. Так как знаки $\pm$ чередуются, то соответствующее соотношение будет

$$z_k = y_{k-1} + 2y_k + y_{k+1}; \quad (3-5.28)$$

предыдущая схема изменяется следующим образом:

$$-\frac{8}{693}, \frac{8}{315}, -\frac{8}{105}, \frac{8}{15}, \frac{8}{3}, \frac{8}{3}, \frac{8}{15}, -\frac{8}{105}, \frac{8}{315}, \dots$$

теперь интерференция уменьшается как кубы расстояний между двумя пиками и в общем случае становится пренебрежимо малой, если только два пика, которые нужно отделить, не слишком близки друг к другу.

Точное положение максимума, вычисленное с помощью вторых разностей (фактически — вторых сумм, ввиду чередующихся знаков схемы), может быть получено следующим образом. Мы проверяем последовательность y_k и обращаем внимание особенно на правильное чередование знаков $\pm$. В некоторых точках мы отмечаем, что чередование нарушается следованием знаков $++$ или $--$. Подчеркиваем эти неправильности следования и устанавливаем, что пик должен лежать между двумя такими значениями p : $p = k$ и $p = k + 1$. Полагаем

$$p = k + \varepsilon. \quad (3-5.29)$$

Составляем отношение двух последовательных z_p [ср. (28)], соответствующих $p = k$ и $p = k + 1$:

$$q_k = \frac{z_k}{z_{k+1}}. \quad (3-5.30)$$

*) См. § 5 гл. 28.

Но z_k пропорционально

$$-\frac{1}{\varepsilon-1} + \frac{2}{\varepsilon} - \frac{1}{\varepsilon+1} = \frac{-2}{(\varepsilon+1)\varepsilon(\varepsilon-1)},$$

а z_{k+1} пропорционально

$$\frac{1}{\varepsilon} - \frac{2}{\varepsilon+1} + \frac{1}{\varepsilon+2} = \frac{2}{(\varepsilon+2)(\varepsilon+1)\varepsilon}.$$

Составляя отношение, получаем

$$q_k = \frac{z_k}{z_{k+1}} = -\frac{\varepsilon+2}{\varepsilon-1},$$

откуда

$$\varepsilon = -\frac{2-q_k}{1+q_k}. \quad (3-5.31)$$

Наконец

$$\theta = \frac{\pi}{N} (k + \varepsilon) \quad (3-5.32)$$

и

$$\left. \begin{aligned} \lambda &= \sin^2 \frac{\theta}{2} = \frac{1 - \cos \theta}{2} && [\text{случай (9)}], \\ \lambda &= \cos \theta && [\text{случай (8)}]. \end{aligned} \right\} \quad (3-5.33)$$

Возвращаясь к собственным значениям первоначальной некалиброванной матрицы A , получаем

$$\lambda_i^* = s\lambda_i. \quad (3-5.34)$$

Наиболее поразительной чертой этого алгоритма является то, что он совершенно свободен от опасного накопления ошибок округления. В то время как в p , q -алгоритме (см. § 3) ошибки округления быстро накапливаются и должны компенсироваться постоянным процессом повторной ортогонализации, спектроскопический алгоритм допускает образование сотен, а быть может даже тысяч итераций без особой опасности искажения со стороны ошибок округления. Так как «сигнал» возрастает пропорционально числу итераций, то возрастание ошибок округления не более быстрое, чем линейное, оставляло бы неизменным «отношение сигнала к помехе».

Опыты с матрицами невысокого порядка не могли обнаружить какого-либо искажения даже после 2000 итераций. Опытов с матрицами высокого порядка еще нет; недостаточно исследовано также статистическое влияние помех, связанных с этим алгоритмом. Однако имеются основания ожидать, что большая точность, которую можно получить с помощью этого метода при независимом определении собственных значений, и высокая разрешающая способность метода при

разделении близких собственных значений не пострадает при увеличении порядка матриц, к которым он применяется.

Целесообразно выбирать число итераций N кратным 180, так как в тригонометрических таблицах используется деление полуокружности на 180 частей. Если, например, мы берем $N=720$, то мы можем исследовать полуокружность по частям в 15° . Так как метод вторых сумм дает хорошие результаты, если два соседних пика отстоят по крайней мере на 4 единицы, то мы можем получить с большой точностью положение какого-нибудь собственного значения на окружности, если он отстоит от своего соседа по меньшей мере на 1° .

6. Построение собственных векторов

В предыдущем параграфе была использована лишь одна составляющая векторов b_k . Метод выделения определенной частоты есть резонансный метод: мы систематически наращиваем величину небольшой периодической составляющей, частота которой как раз совпадает с частотой вынужденных колебаний. В предыдущем параграфе нас интересовало лишь точное *положение* максимума. Но если нашей целью является получение собственного *вектора*, соответствующего некоторому собственному значению, то для нас представляет интерес также и *величина* пикового значения. Метод «вторых сумм» [ср. (5.28)] является ценным орудием также для изолирования пика от затемняющего влияния соседних пиков. Если мы нашли, что максимум лежит между $p=k$ и $p=k+1$ при бóльшей амплитуде в точке $p=k$ (или между $p=k$ и $p=k-1$ при бóльшей амплитуде в точке $p=k$), то величина

$$z_k = y_{k-1} + 2y_k + y_{k+1} \quad (3-6.1)$$

будет пропорциональна одной из составляющих собственного вектора u_j . Для того чтобы получить весь собственный вектор u_j , мы должны повторить тот же расчет для каждой составляющей.

Эти операции можно выполнить чрезвычайно экономично. Пусть координаты векторов b_α уже вычислены и сохраняются в запоминающем устройстве вычислительной машины. Умножим каждый из этих последовательных векторов на заранее установленный множитель веса ρ_i и образуем сумму, получая таким образом

$$\tilde{u}_j = b_0 + \sum_{\alpha=1}^{N-1} \rho_\alpha b_\alpha, \quad (3-6.2)$$

где веса ρ_α определяются по формуле

$$\rho_\alpha = \left(1 + \cos \frac{\pi}{N} \alpha\right) \cos \frac{\pi}{N} k\alpha. \quad (3-6.3)$$

Выписывается лишь конечная сумма $\bar{u}_j^*$). Длина вектора не нормирована. (Черточка над $\bar{u}_j$ отмечает тот факт, что это не точный собственный вектор u_j , а только близкое к точному его приближенное значение.) Искажающим влиянием со стороны других собственных векторов можно пренебречь, если пик близ $p = k$ и ближайший к нему пик отстоят друг от друга по меньшей мере на 4 единицы.

С помощью этого метода можно построить любой отдельный собственный вектор матрицы A или все ее векторы. Если A симметрична, то полученные таким образом векторы $\bar{u}_j$ можно проверить на ортогональность. Если же A несимметрична, то надо одновременно оперировать и над матрицей $\tilde{A}$, повторив весь процесс с другим пробным вектором $\tilde{b}_0$, с заменой матрицы A матрицей $\tilde{A}$. Векторы $\tilde{b}_\alpha$ снова передаются в запоминающее устройство, но взвешивание (2) производится опять по-прежнему. Результирующая сумма снова выписывается и дает вектор $\bar{v}_j$. Полученные векторы $\bar{u}_j$ и $\bar{v}_j$ должны удовлетворять соотношениям биортогональности (2-6.16).

7. Итерационное решение обширных линейных систем

Применение чебышевских полиномов к задаче нахождения собственных значений матриц с вещественными собственными значениями приводит к методу решения обширных линейных систем путем последовательных итераций, минуя фактическое обращение матриц. На первых порах мы встречаемся при этом с тем затруднением, что чебышевские полиномы применимы непосредственно только в *вещественной* области, тогда как собственные значения произвольной несимметрической матрицы A являются, вообще говоря, комплексными числами. Это затруднение можно преодолеть двояким путем. Один из них состоит в том, что заданное уравнение

$$Ax = b \quad (3-7.1)$$

мы преобразуем по методу наименьших квадратов в уравнение

$$\tilde{A}^* Ax = \tilde{A}^* b. \quad (3-7.2)$$

Новая матрица $\tilde{A}^* A$ «положительно определена», т. е. ее собственные значения все вещественны и даже положительны **). Если первоначальная матрица A была надлежаще калибрована по строкам и столбцам

*) В формуле (3) постоянное целое k есть то значение p , ближайшее к максимуму функции $F(p)$ (5.18), на основе которого вычислено собственное значение λ_j , соответствующее искомому собственному вектору u_j .

По формуле (2) вычисляется последовательно каждая координата вектора u_j в отдельности и таким путем получается весь вектор.

**) Здесь и дальше в этом параграфе предполагается, что матрица A не вырождается, т. е. что ни одно из ее собственных чисел не равно нулю.

(ср. II, 26), то диагональные элементы в $\tilde{A}^*A$ будут близки к 1, а недиагональные элементы будут лежать между -1 и 1 .

«Симметризация матрицы A » путем умножения ее слева на $\tilde{A}^*$ есть простая, но трудоемкая операция. Она эквивалентна n итерациям, и если n велико, то эта операция совсем нетривиальна. Кроме того, если матрица A сама содержит много нулей, то относительно матрицы $\tilde{A}^*A$ это уже не будет верно, и таким образом может потеряться большое практическое преимущество первоначальной матрицы A . Поэтому иногда предпочтительно не составлять явно матрицы $\tilde{A}^*A$, а проводить необходимую для итераций операцию $b_1 = A^*Ab_0$ в два приема, а именно так:

$$b'_1 = Ab_0, \quad b_1 = \tilde{A}^*b'_1. \quad (3-7.3)$$

Однако существует совершенно иной путь приведения A к симметрическому виду, при котором можно избежать умножения ее слева на $\tilde{A}^*$; это удастся сделать за счет удвоения размера матрицы. Заданную линейную систему уравнений мы расширяем следующим образом:

$$Ax = b, \quad \tilde{A}^*y = 0 \quad (3-7.4)$$

и рассматриваем

$$z = (y, x) \quad (3-7.5)$$

как вектор с $2n$ координатами. На первый взгляд добавление второго уравнения кажется ненужным придатком, так как оно имеет только тривиальное решение $y = 0$. Однако расширенная матрица

$$B = \begin{bmatrix} 0 & A \\ \tilde{A}^* & 0 \end{bmatrix} \quad (3-7.6)$$

обладает тем большим преимуществом, что она всегда эрмитова:

$$\tilde{B}^* = B. \quad (3-7.7)$$

Следовательно, собственные числа матрицы B всегда вещественны. Применение метода Гершгорна [17] к этой новой матрице приводит к интересным заключениям. В § 5 мы нашли, что можно найти верхнюю границу s собственных чисел матрицы A , рассмотрев сумму абсолютных величин элементов каждой строки и выбрав наибольшее из этих n чисел, либо — сумму абсолютных величин элементов каждого столбца и выбрав наибольшее из этих n чисел. Наименьшее из этих двух найденных значений может быть выбрано в качестве верхнего предела для s . В рассматриваемом случае следует взять в качестве s для расширенной матрицы (6) наибольшее из этих двух чисел. Положим теперь

$$C = \frac{2}{s} B. \quad (3-7.8)$$

Собственные значения матрицы C будут тогда лежать между -2 и $+2$.

Но то же самое s можно использовать и для другой цели. Задача нахождения собственных значений матрицы B приводит к равенствам

$$Ax = \rho y, \quad \tilde{A}^* y = \rho x, \quad (3-7.9)$$

что дает

$$\tilde{A}^* Ax = \rho^2 x. \quad (3-7.10)$$

Отсюда видно, что $2n$ собственных значений матрицы B являются *квадратными корнями* $\pm \sqrt{\lambda}$ из собственных значений матрицы $\tilde{A}^* A$. Так как наибольшее по абсолютной величине собственное значение матрицы B , как мы видели, не превышает s , то мы теперь получаем

$$\lambda_M \leq s^2 \quad (3-7.11)$$

в качестве оценки верхней границы собственных значений матрицы $\tilde{A}^* A$, не вычисляя фактически ее элементов. Кроме того, мы приходим к выводу, что собственные значения матрицы

$$S_0 = \frac{1}{s^2} \tilde{A}^* A \quad (3-7.12)$$

лежат между 0 и 1.

Для последующих рассуждений примем, что мы пришли к симметрической матрице путем умножения A слева на $\tilde{A}^*$, а не «методом расширения» (4), хотя алгоритм, который мы получим в результате наших рассуждений, легко может быть истолкован и для случая (4).

Нашу задачу мы приведем теперь к задаче нахождения собственных значений (ее решение было указано в § 5). С этой целью запишем заданное уравнение (1) в форме

$$S_0 x + c = 0, \quad (3-7.13)$$

где

$$c = -\frac{\tilde{A}^* b}{s^2}. \quad (3-7.14)$$

Во многих задачах, возникающих при решении самосопряженных дифференциальных уравнений, встречаются матрицы, которые имеют только положительные собственные значения. В этом случае нет нужды в какой-либо симметризации, и мы можем прямо положить

$$S_0 = \frac{1}{s} A, \quad c = \frac{1}{s} b. \quad (3-7.15)$$

Наша линейная система уравнений с n неизвестными

$$\left. \begin{aligned} s_{11}x_1 + s_{12}x_2 + \dots + s_{1n}x_n + c_1 &= 0, \\ \vdots & \\ s_{n1}x_1 + s_{n2}x_2 + \dots + s_{nn}x_n + c_n &= 0. \end{aligned} \right\}$$

ной процедурой мы расставим составляющие последовательных собственных векторов в последовательные столбцы матрицы. В нашей задаче, по условию, взаимная ортогональность собственных векторов матрицы S_0 обеспечена. Эти векторы обозначены через $u_1, u_2, \dots, u_n$. Дополнительная $(n+1)$ -я координата указана отдельно:

$$\begin{array}{c|c|c|c|c} \lambda=0 & \lambda_1 & \lambda_2 & \dots & \lambda_n \\ \hline x & u_1 & u_2 & \dots & u_n \\ 1 & 0 & 0 & \dots & 0 \end{array} \quad \begin{array}{c|c|c|c|c} \tilde{\lambda} & \lambda_1 & \lambda_2 & \dots & \lambda_n \\ \hline 0 & u_1 & u_2 & \dots & u_n \\ 1 & \frac{u_1 c}{\lambda_1} & \frac{u_2 c}{\lambda_2} & \dots & \frac{u_n c}{\lambda_n} \end{array} \quad (3-7.19)$$

Начнем с пробного вектора

$$b_0 = (0, 0, \dots, 0, 1). \quad (3-7.20)$$

Разлагаем его по главным осям матрицы S :

$$b_0 = \beta_0 u_0 + \beta_1 u_1 + \dots + \beta_n u_n. \quad (3-7.21)$$

Коэффициенты этого разложения β_i получаются в результате скалярного умножения вектора b_0 на соответствующие собственные векторы сопряженной системы; следовательно,

$$\beta_0 = 1, \quad \beta_i = \frac{u_i c}{\lambda_i}. \quad (3-7.22)$$

Станем действовать на b_0 некоторым надлежаще выбранным полиномом $P_m(S)$. Согласно формуле (2.13) получаем

$$P_m(S) b_0 = P_m(0) u_0 + P_m(\lambda_1) \beta_1 u_1 + \dots + P_m(\lambda_n) \beta_n u_n. \quad (3-7.23)$$

Нашей целью является выделение *одного* вектора u_0 без какого-либо *искажения* со стороны других векторов u_i . Поэтому мы постараемся найти такой полином $P_m(\lambda)$, который принимал бы значение 1 при $\lambda=0$ и оставался бы малым при всех значениях λ , заключенных между 0 и 1. Хотя значение полинома не может сразу упасть с 1 до 0, но мы можем построить полином, значения которого будут резко снижаться от начального значения $y=1$ при $\lambda=0$ и затем оставаться все время малыми. Полиномы Чебышева второго рода пригодны для решения этой задачи.

Мы строим последовательность векторов $b_1, b_2, \dots, b_n$ с помощью следующего простого правила. Образует матрицу

$$C = 2I - 4S \quad (3-7.23')$$

и составляем векторы b_{k+1} по единообразной рекуррентной схеме, которая на каждом этапе зависит от двух векторов, а именно последнего полученного вектора b_k и предпоследнего вектора b_{k-1} :

$$b_{k+1} = Cb_k - b_{k-1}. \quad (3-7.24)$$

Схема начинается с двух векторов:

$$b_1 = b_0, \quad b_2 = Cb_1.$$

Затем уже мы используем правила (24); получаем

$$b_3 = Cb_2 - b_1$$

и т. д. Наконец, достигнув некоторого b_m , мы делим этот вектор на m . Полиномиальный оператор, полученный таким образом, можно записать в виде

$$P_m(\lambda) = \frac{\sin m\theta}{m \sin \theta} \quad \left(\lambda = \sin^2 \frac{\theta}{2} \right). \quad (3-7.25)$$

Эта функция при малых θ имеет характер некоторой универсальной функции от одной переменной ξ , если определить ξ равенством

$$\xi = 4m^2\lambda. \quad (3-7.26)$$

Многочлен $P_m(\lambda)$, рассматриваемый как функция от ξ , хорошо аппроксимируется функцией

$$\Phi(\xi) = \frac{\sin \sqrt{\xi}}{\sqrt{\xi}}. \quad (3-7.27)$$

Эта функция при $\xi = 0$ принимает значение $\Phi = 1$; далее она резко убывает с возрастанием ξ , но дает вторичные максимумы и минимумы

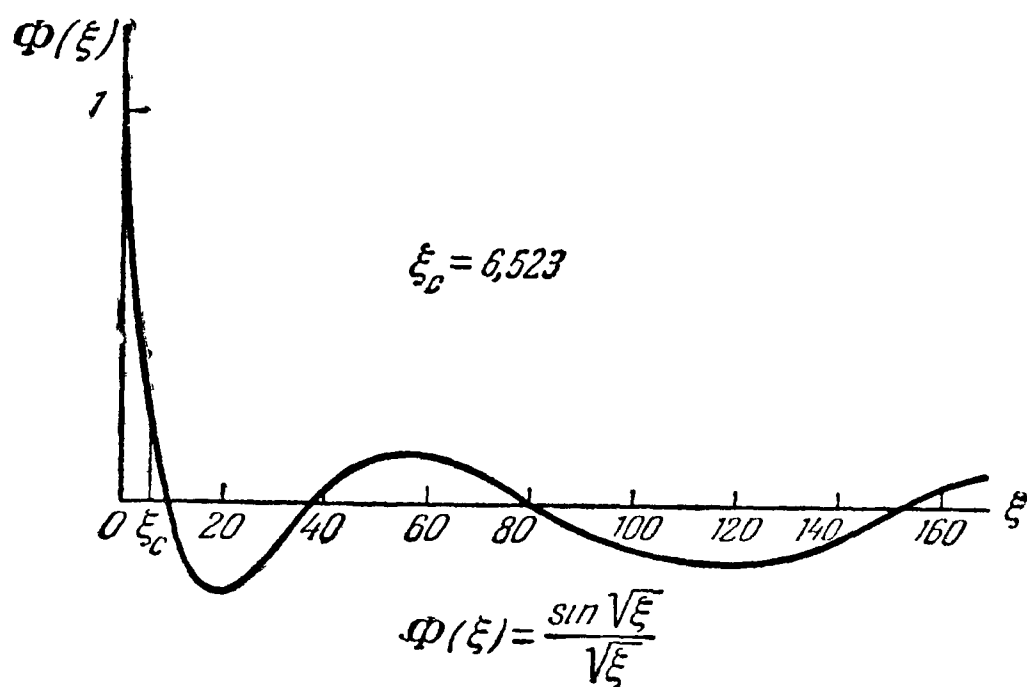


Рис. 23.

(рис. 23). Когда m , т. е. число итераций, возрастает, то для определенного собственного значения λ_i величина ξ , согласно формуле (26), возрастает со скоростью m^2 . Увеличивая число итераций, мы даже при малом λ придем в конце концов к большим значениям ξ , при которых $\Phi(\xi)$ уже очень мало. Однако этот процесс очень медленно сходится, если исходная линейная система векторов весьма косоугольна. Для того чтобы избежать чрезмерно большого числа итераций, предпо-

читательно закончить нашу рекуррентную процедуру после m этапов и затем повторить весь цикл вторично, а возможно, и в третий, и четвертый раз. Это повторение может быть без труда выполнено, если воспользоваться следующим небольшим видоизменением нашего приема.

Целесообразно выбрать число m , т. е. длину каждого цикла, равным некоторой степени числа 2, например, положить $m = 128 = 2^7$. Тогда деление на m осуществляется простым отделением 7 двоичных разрядов. Таким образом, мы получаем

$$b_1, b_2, \dots, b_m; b'_m = b_m / m$$

и продолжаем наш прежний прием, приняв b'_m в качестве исходного вектора b_{m+1} второго цикла. Отклонение от рекуррентного процесса (24) происходит лишь на первом этапе, а именно при образовании вектора b_{m+2} по формуле

$$b_{m+2} = Cb_{m+1}, \quad (3-7.28)$$

когда, следовательно, не производится вычитания вектора b_m . Но сразу же после этого мы опять возвращаемся к рекуррентной формуле

$$b_{m+3} = Cb_{m+2} - b_{m+1}$$

и продолжаем рекуррентный процесс без изменений, пока не доходим до вектора

$$b'_{2m} = b_{2m} / m.$$

Затем начинается третий цикл. Таким образом осуществляется ν циклов. В результате мы получаем полиномный оператор $Q(\lambda)$, определяемый формулой

$$Q_{\nu m}(\lambda) = [P_m(\lambda)]^\nu. \quad (3-7.29)$$

Рассмотрим, в какой связи находится точность полученного решения с числом ν выполненных циклов и длиной m каждого из них. С этой целью подставим последний вектор $b'_{\nu m}$ в заданное уравнение (17) и образуем остаток. Введем обозначение

$$\nu m = N;$$

тогда

$$\begin{aligned} r = Sb_N &= \beta_1 \lambda_1 Q_N(\lambda_1) u_1 + \beta_2 \lambda_2 Q_N(\lambda_2) u_2 + \dots + \beta_n \lambda_n Q_N(\lambda_n) u_n = \\ &= (c \cdot u_1) Q_N(\lambda_1) u_1 + (c \cdot u_2) Q_N(\lambda_2) u_2 + \dots + (c \cdot u_n) Q_N(\lambda_n) u_n. \end{aligned} \quad (3-7.30)$$

Но $(c \cdot u_1)$, $(c \cdot u_2)$, ..., $(c \cdot u_n)$ суть составляющие вектора c по направлениям главных осей матрицы S_0 . Рассматривая снова наше решение как решение неоднородного уравнения

$$S_0 x = -c$$

и полагая

$$x = b_N + y, \quad (3-7.31)$$

получаем для поправочного члена y равенство

$$S_0 y = -r. \quad (3-7.32)$$

Даже не вычисляя y , можно теоретически предсказать, насколько точным будет решение b_N . Векторы $u_1, u_2, \dots, u_n$ ортогональны друг другу. Если принять, что величины λ_i расположены в возрастающем порядке, то наименьшим собственным числом будет λ_1 ; наибольшая ошибка будет в направлении соответствующего собственного вектора. Та составляющая вектора в этом направлении, которую мы отбросили, имеет множитель $Q_N(\lambda_1)$. Допустим теперь, что наименьшее собственное значение λ_1 больше некоторой критической величины λ_c , определенной равенством

$$\lambda_c = \frac{\xi_c}{4m^2}, \quad (3-7.33)$$

где

$$\xi_c = (2,554)^2 = 6,523. \quad (3-7.34)$$

Это частное значение для ξ выбрано в результате учета первого вторичного минимума функции $(\sin x)/x$, которого она достигает в точке $x = 4,4934$. Значение этого минимума есть $-0,2172$. Той же величины с положительным знаком достигает эта функция и раньше в точке $x = 2,5536$. За этой точкой функция $\Phi(\xi)$ никогда не подымается выше $0,22$. Следовательно, мы можем сказать, что если наименьшее собственное значение λ_1 нашей матрицы S_0 не опускается ниже некоторой критической величины

$$\lambda_1 \geq \left(\frac{1,277}{m} \right)^2, \quad (3-7.35)$$

то относительная ошибка нашего решения есть

$$\eta \leq (0,22)^n. \quad (3-7.36)$$

Эта «относительная ошибка» означает следующее. Мы не можем оценить точность решения линейной системы путем сравнения вычисленного значения некоторой одной составляющей с ее истинным значением. Истинное значение отклонения некоторой составляющей от ее приближения x_i случайно может быть весьма мало и даже равняться нулю, тогда как отклонения для других составляющих x_j могут быть весьма велики. Правильной мерой точности служит длина поправочного вектора, поделенная на длину истинного вектора, т. е. корень квадратный из частного от деления суммы квадратов отклонений каждой составляющей на сумму квадратов всех неизвестных составляющих:

$$\eta = \sqrt{\frac{(\delta x_1)^2 + (\delta x_2)^2 + \dots + (\delta x_n)^2}{x_1^2 + x_2^2 + \dots + x_n^2}}. \quad (3-7.37)$$

Здесь δx_i означает разность между истинным значением x_i и вычисленным значением x'_i .

Соотношение (36) показывает, что *число повторенных циклов является определяющим для точности нашего решения*. Два цикла дадут нам точность до $4,72\%$, три цикла — точность до $1,02\%$, а четыре цикла — точность до $0,22\%$. Это — оценочные точности, а оценка точности основывается на худших возможностях. Фактическая ошибка может быть гораздо меньше, ибо собственные значения λ_i нашей матрицы могут определять (по формулам (26) и (27)) такие значения для $\Phi(\xi)$, которые ближе к нулевым, чем к максимальным значениям графика, изображенного на рис. 23.

С другой стороны, формула (35) показывает, что длина каждого цикла определяет разрешающую способность метода. В случае очень малого λ_1 минимальное число итераций, необходимых для того, чтобы успешно отделить правильное решение от наслоений со стороны наименьшего собственного вектора, может стать очень большим. Мы видели, однако, что системы физического происхождения, правые части которых суть результаты измерений, едва ли допускают снижение значений λ_1 ниже 10^{-4} (иначе помехи при измерениях заглушат математически правильное решение). Но тогда цикл из $m = 128 = 2^7$ итераций будет достаточен для того, чтобы очистить решение от всех искажений, вызванных собственными векторами матрицы S_0 . Кроме того, точность решения едва ли должна превышать 5% ввиду неточностей физических наблюдений. Мы можем, таким образом, сказать, что двойной цикл по 128 итераций будет достаточен практически для всех систем физического происхождения независимо от размера матрицы¹⁾.

В дальнейшем для краткости мы присвоим схеме (24) образования последовательности векторов название «С-итерации»*).

8. Остаточное испытание

Если мы произвели полный анализ собственных значений матрицы S_0 , то по длине циклов в вышеописанном алгоритме и по числу использованных циклов мы можем заранее судить о том, какова будет минимальная точность (максимальная ошибка) нашего решения. Фактическая точность может оказаться гораздо более благоприятной.

¹⁾ Полином (25) имеет сильный максимум не только при $\theta = 0$, но и при $\theta = \pi$. Это значит, что выдвигаются на первый план окрестность $\lambda = 0$ и окрестность $\lambda = 1$. Поэтому спектр собственных значений не должен простирается вплоть до $\lambda = 1$. Оценка Гершгорна обычно столь сильно преувеличивает наибольшее собственное значение, что опасность того, что матрица S_0 имеет собственное значение $\lambda = 1$, весьма мала. Однако чтобы полностью исключить такую возможность, мы можем выбрать в качестве матрицы S_0 матрицу $\tilde{A}A$, разделенную на $1,05 s^2$.

*) Обратим внимание на итеративный метод решения систем линейных уравнений, указанный в [9], а также и на дополнения к нему, изложенные в [10] и [11]. Укажем также и методы, изложенные в [12], [12a] и [13] и на статьи [14], [15] и [16].

Поэтому было бы хорошо установить, насколько близко наше решение к математически совершенному решению. На первый взгляд может показаться, что наиболее естественный метод определения точности решения заключается в подстановке предполагаемого решения в заданное уравнение: если окажется, что заданное уравнение удовлетворяется, то ответ верен.

Однако этот метод таит в себе большие опасности. Правда, мы можем таким путем убедиться в правильности абсолютно точного решения: если при подстановке в выражение $S_0 x + c$ найденного решения остаточный вектор

$$r = S_0 x + c \quad (3-8.1)$$

оказывается нулем, то мы считаем, что нашли надлежащий вектор x . Однако такие системы встречаются очень редко; причем обычно это бывает лишь в особо простых, но не характерных случаях, когда элементами матрицы S_0 служат небольшие целые числа. В большинстве случаев ограниченная точность вычислений исключает возможность точного ответа. Поэтому не следует ожидать, что остаточный вектор r , полученный в результате подстановки найденного решения в заданное уравнение, окажется точно нулем, можно лишь ожидать, что он будет весьма малым. И все же одним из парадоксов линейных систем является то, что это простое испытание — мы можем назвать его «остаточным испытанием» — необязательно будет правильной мерой точности нашего решения. Хотя нулевой остаточный вектор свидетельствует о совершенно точном ответе, малый остаточный вектор *необязательно* свидетельствует о *хорошем приближенном ответе*.

Следующий пример может служить иллюстрацией этого обстоятельства. Допустим, что наибольшее собственное значение матрицы S_0 есть $\lambda_0 = 1$, а наименьшее $\lambda_n = 10^{-4}$ и что истинное решение нашей задачи есть

$$x = u_1 + u_n,$$

т. е. наименьший и наибольший собственные векторы входят в решение с одинаковым весом. Предположим, что нам неизвестны собственные векторы матрицы S_0 и что мы только нашли с помощью некоторого численного алгоритма решение, которое оказалось равным

$$x' = u_1 + 1,01 u_n.$$

Допустим еще, что каким-то другим способом мы нашли второй ответ

$$x'' = 4u_1 + u_n.$$

В первом случае остаточный вектор есть

$$x' - x = 0,01 u_n,$$

во втором случае он равен

$$x'' - x = 3u_1.$$

Следовательно, относительная ошибка в первом случае составляет

$$\frac{0,01}{\sqrt{1+1}} = 0,0071,$$

а во втором случае

$$\frac{3}{\sqrt{2}} = 2,13.$$

То есть в первом случае ошибка в решении составляет менее чем 1° , а во втором — свыше 200° . Остаточное испытание дает в первом случае

$$r' = 0,01 u_n,$$

во втором случае

$$r'' = 0,0003 u_1.$$

Если судить по остаточному испытанию, то следует предпочесть первое решение, ибо ошибка в 33 раза более, чем при втором. Мы охотно, следовательно, примем второе решение как лучшее. И все же первое решение является весьма удовлетворительным, тогда как второе — совершенно негодно. Мы видим, следовательно, что простое остаточное испытание без последующего исследования не может служить мерой точности или неточности данного решения (в рассматриваемом случае ошибка в направлении собственного вектора с малым собственным числом умножается на очень малый множитель и поэтому практически стирается, хотя эта ошибка содержится в самом решении и ощутима, если не выполнено умножение на матрицу S_0).

Однако остаточное испытание можно сделать надежным, если не останавливаться на получении остатка, а использовать последний как новую правую часть уравнения и повторить еще раз всю процедуру. В конце этого приема мы снова составляем остаток. Если этот новый остаток значительно уменьшился сравнительно с первым, то это служит указанием на то, что мы действительно нашли решение данной задачи. Если, наоборот, остаток остался к концу нового цикла без существенного изменения, то это указывает на наличие влияния весьма малого собственного значения, с которым наш процесс итерации не в состоянии справиться.

9. Наименьшее собственное значение эрмитовой матрицы

Итерационный процесс, изложенный в § 7, дает новый подход к задаче малых собственных чисел. Обычно, если нам нужно определить наименьшие собственные числа матрицы, мы сначала обращаем эту матрицу и затем отыскиваем наибольшие собственные значения. С этой целью мы образуем векторы [ср. (3.8)]

$$b_0, A^{-1}b_0, A^{-2}b_0, \dots \quad (3-9.1)$$

и соответствующие векторы сопряженной системы. Но уравнение

$b_1 = A^{-1}b_0$ эквивалентно уравнению

$$Ab_1 = b_0, \quad (3-9.2)$$

и так как мы можем решить это уравнение путем последовательных итераций, то обращения матрицы больше не требуется. Далее, после того, как мы нашли b_1 , мы подставляем b_0 на место b_1 и повторяем ту же процедуру. Таким образом, мы получаем решение уравнения

$$Ab_2 = b_1. \quad (3-9.3)$$

Правда, мы описали выше итерационное решение системы линейных уравнений лишь для случая положительно определенных матриц. Но матрицы, встречающиеся в задачах теории колебаний, обычно принадлежат к типу положительно определенных матриц. Кроме того, если A есть произвольная, вообще говоря, комплексная матрица, наименьшие собственные значения которой и нужно определить, то эта задача собственных значений тесно связана с сопряженной задачей, приводящей к эрмитовой матрице

$$S_0 = \tilde{A}^* A. \quad (3-9.4)$$

Хотя сама матрица A может иметь и комплексные элементы, матрица S_0 имеет всегда вещественные собственные числа, так как она удовлетворяет эрмитову условию

$$\tilde{S}_0^* = S_0 \quad (3-9.5)$$

[ср. (2-8.29)]. Кроме того, в силу тех свойств матрицы (4), которые вытекают из ее связи с методом наименьших квадратов, собственные числа матрицы S_0 не только вещественны, но и положительны. Поэтому уравнение

$$S_0 b_1 = b_0 \quad (3-9.6)$$

всегда может быть решено методом S -итераций, рассмотренным в § 7.

В области весьма малых собственных чисел нам не удалось решить задачу обращения. Это, однако, не вызывает серьезных трудностей. Главной целью обращения является нахождение обратных величин собственных чисел, т. е. превращение малых собственных чисел в большие, и больших в малые. Наш процесс дает возможность достигнуть этой цели. С помощью двух циклов из m итераций мы образовали следующую функцию от собственного значения λ :

$$4m^2 \frac{\xi - (\sin \sqrt{\xi})^2}{\xi^2} \quad (\xi = 4m^2\lambda). \quad (3-9.7)$$

Эта функция отличается от λ^{-1} только в области малых ξ , т. е. в области весьма малых λ . Но приравнивание λ^{-1} функции (7) остается законным даже в этой области. Образует скаляры

$$c_0 = b_0 b_0^*, \quad c_1 = b_0 b_1^*, \quad c_2 = b_0 b_2^* = b_1 b_1^*, \quad c_3 = b_1 b_2^*. \quad (3-9.8)$$

Эти скаляры остаются вещественными даже для комплексных векторов b_i . Квадратное уравнение

$$\begin{vmatrix} 1 & \mu & \mu^2 \\ c_0 & c_1 & c_2 \\ c_1 & c_2 & c_3 \end{vmatrix} = 0 \quad (3-9.9)$$

определяет два самых больших собственных значения «обращенной» матрицы*). Чтобы получить наименьшее собственное значение матрицы A , следует взять *большой* из двух корней: $\mu_1 > \mu_2$. Соответствующий собственный вектор u_1 определяется тогда равенством

$$u_1 = b_2 - \mu_2 b_1. \quad (3-9.10)$$

Этот метод действует еще лучше, если начинать с такого пробного вектора, который выдвигает вперед наименьшее собственное значение. Этого мы можем достичь, если применим два-три двойных цикла S -итераций к случайному вектору r_0 . Пусть полученный таким образом вектор будет w_0 ; применим к нему еще раз двойной цикл итераций — полученный вектор w_1 можно считать хорошим приближением собственного вектора u_1 , и тогда отношение

$$\mu = \frac{w_0 w_1^*}{w_0 w_0^*} \quad (3-9.11)$$

дает преобразованное значение собственного числа λ .

Обратное преобразование от μ к λ можно выполнить следующим образом. Сначала вычисляем

$$\frac{2m}{\pi \sqrt{\mu}} = v. \quad (3-9.12)$$

Затем получаем u , решая уравнение

$$\frac{u}{\sqrt{1 - \left(\frac{\sin \pi u}{\pi u}\right)^2}} = v. \quad (3-9.13)$$

Для $v \leq 5$ это решение можно получить с помощью табл. III (стр. 497). Для значений $v > 5$ можно с достаточной точностью положить

$$u = v \sqrt{1 - \left(\frac{\sin \pi v}{\pi v}\right)^2}. \quad (3-9.14)$$

Найдя u , мы получим λ при помощи соотношения

$$\lambda = \left(\frac{\pi}{2m} u\right)^2. \quad (3-9.15)$$

*) См. § 3.

10. Собственное значение произвольной матрицы, отличное от наибольшего

Задача нахождения не наибольшего собственного значения произвольной матрицы имеет весьма важное значение. Может случиться, что нас интересует какое-либо одно частное собственное значение матрицы A , которое не является ни наибольшим, ни наименьшим. Допустим, что мы имеем в своем распоряжении довольно хорошее приближение λ_0 для λ . Нам, следовательно, известно, что искомое λ отличается от λ_0 лишь на небольшую величину ε . Задача состоит в том, чтобы найти ε .

Выделение одного частного собственного значения представляет задачу нелегкую, так как в обычной итерационной технике неизбежно будут превалировать наибольшие собственные значения над искомым собственным значением. Следующий метод свободен от этого затруднения. Так как

$$\lambda = \lambda_0 + \varepsilon, \quad (3-10.1)$$

то матрица $A - \lambda_0 I - \varepsilon I$ имеет нулевое собственное значение. А это значит, что матрица

$$A_0 = A - \lambda_0 I \quad (3-10.2)$$

имеет малое собственное значение ε . Это ε будет, вообще говоря, комплексным, если даже заданное λ_0 есть вещественное число. Нам нужно найти квадратичное приближение для ε .

Мы снова будем иметь в виду обратную матрицу A_0^{-1} и постараемся найти ее наибольшее собственное значение методом § 9. Для этого, однако, нам надо начать с такого пробного вектора b_0 , который достаточно мало отличается от вектора, соответствующего наибольшему собственному значению. Мы рассуждаем теперь следующим образом. Если пренебречь числом ε и считать, что A_0 имеет собственное значение нуль, то симметризованная эрмитова матрица $S_0 = \tilde{A}_0^* A_0$ также имеет нулевое собственное значение. Кроме того, собственный вектор, соответствующий нулевому собственному значению, *одинаков* для обеих матриц. Но это значит, что даже если ε не равно нулю, но мало, то собственный вектор x_0 матрицы S_0 , соответствующий наименьшему собственному значению, будет иметь большую составляющую в направлении искомого собственного вектора матрицы A_0 .

Мы находим поэтому, пользуясь методом предыдущего параграфа, наименьший собственный вектор *) эрмитовой матрицы S_0 :

$$S_0 x_0 = \rho_0 x_0. \quad (3-10.3)$$

*) То есть собственный вектор, соответствующий наименьшему собственному числу.

С этой целью мы либо улучшаем произвольный случайный вектор путем нескольких S -итераций и рассматриваем непосредственно результирующий вектор в качестве нашего x_0 , либо пользуемся более точным квадратичным приближением (9.9). Во всяком случае мы теперь имеем начальный вектор $b_0 = x_0$. Нам нужен еще подобный же начальный вектор для $\tilde{A}_0$. Теперь симметризованная матрица будет другая, а именно

$$S'_0 = A_0^* \tilde{A}_0. \quad (3-10.4)$$

Но если уравнение (3) умножить слева на A_0 , то получим

$$A_0 \tilde{A}_0^* A_0 x_0 = \rho_0 A_0 x_0, \quad (3-10.5)$$

а заменив i на $-i$:

$$(A_0^* \tilde{A}_0) A_0^* x_0^* = \rho_0 A_0^* x_0^*. \quad (3-10.6)$$

Поэтому

$$\tilde{x}_0 = A_0^* x_0^*. \quad (3-10.7)$$

Мы имеем теперь *пару* начальных векторов. Если ρ_0 весьма мало, то линейная аппроксимация будет достаточной для получения ε . В этом случае мы непосредственно получаем

$$\varepsilon = \frac{x_0 A \tilde{x}_0}{x_0 \tilde{x}_0} = \rho_0 \frac{x_0 x_0^*}{x_0 A^* x_0^*}. \quad (3-10.8)$$

Но если ρ_0 не очень мало, то мы приступаем к построению x_1 на основе равенства $Ax_1 = x_0$, которое означает, что

$$S_0 x_1 = \tilde{A}_0^* x_0. \quad (3-10.9)$$

Это уравнение решается с помощью предыдущего итерационного приема. С этой целью мы берем N достаточно большим, чтобы перевести ρ_0 в безопасную область: $\rho_0 \geq \lambda_c$ [ср. (7.33)]. Затем мы продолжаем следующим образом: рассматриваем наши пробные векторы x_0 , $\tilde{x}_0$ не в виде b_0 , $\tilde{b}_0$, а как среднюю пару b_1 , $\tilde{b}_1$. Тогда $b_2 = x_1$, в то время как

$$\left. \begin{aligned} b_0 &= A_0 x_0 = \tilde{x}_0^*, \\ \tilde{b}_0 &= \tilde{A}_0 \tilde{x}_0. \end{aligned} \right\} \quad (3-10.10)$$

(Последний вектор можно заменить вектором $\rho_0 x_0^*$, если (3) точно удовлетворяется. Однако мы не будем рассчитывать на точное выполнение равенства (3).) Теперь мы имеем пять векторов

$$(\tilde{x}_0^* b_0) \quad (x_0, \tilde{x}_0) \quad x_1 \quad (3-10.11)$$

и можем построить четыре основных скаляра c_i :

$$\left. \begin{aligned} c_0 &= \tilde{x}_0^* \tilde{b}_0, & \gamma_1 &= \frac{\tilde{x}_0^* \tilde{b}_0}{\tilde{x}_0^* \tilde{x}_0}, \\ c_1 &= x_0 \tilde{b}_0 = \tilde{x}_0 \tilde{x}_0^*, & \gamma_2 &= \frac{x_0 \tilde{x}_0}{\tilde{x}_0^* \tilde{x}_0}, \\ c_2 &= x_0 \tilde{x}_0 = \tilde{b}_0 x_1, & \gamma_3 &= \frac{x_1 \tilde{x}_0}{\tilde{x}_0^* \tilde{x}_0}, \\ c_3 &= x_1 \tilde{x}_0, \end{aligned} \right\} \quad (3-10.12)$$

вместе с квадратным уравнением

$$\begin{vmatrix} \varepsilon^2 & \varepsilon & 1 \\ \gamma_1 & 1 & \gamma_2 \\ 1 & \gamma_2 & \gamma_3 \end{vmatrix} = 0. \quad (3-10.13)$$

Членами, содержащими γ_1 , часто можно пренебречь ввиду их малости, и в этом случае наше уравнение приводится к виду

$$\varepsilon^2 (\gamma_3 - \gamma_2^2) + \varepsilon \gamma_2 - 1 = 0. \quad (3-10.14)$$

Таким образом, мы получаем поправку для второго приближения, которую нужно прибавить к предварительному приближенному значению λ_0 собственного значения λ^1).

ЛИТЕРАТУРА К ГЛАВЕ III

- [1] Horvay G., Solution of Large Equations Systems and Eigenvalue Problems by Lanczos' Matrix Iteration Method (General Electric Company, Report No. KAPL-1004, Technical Information Service, Oak Ridge, Tenn., 1953).
- [1a] Фаддеев Д. К. и Фаддеева В. Н., Вычислительные методы линейной алгебры, Физматгиз, 1960.
- [2] Householder A. S., Forsythe G. S. and Germond H. H., Monte Carlo Methods (Nat. Bur. Standards, Washington, D. C., 1951, Applied Math. Series 12).

Статьи:

- [3] Hestenes M. R. and Karush W., A Method of Gradients for the Calculation of the Characteristic Matrix, I. Research, Nat. Bur. Standards 47, 45 (1951).
- [4] Hestenes M. R. and Stiefel E., Method of Conjugate Gradients for Solving Linear Systems, I. Research, Nat. Bur. Standards 49, 409 (1952).
- [5] Lanczos C., An Iteration Method for the Solution of the Eigenvalue Problem of Linear Differential and Integral Operators, I. Research, Nat. Bur. Standards 45, 255 (1951).
- [6] Lanczos C., Solution of Systems of Linear Equations by Minimized Iterations, I. Research, Nat. Bur. Standards 49, 33 (1952).

¹⁾ Интересное приложение этого метода к решению алгебраических уравнений мы вынуждены опустить за недостатком места.

- [7] Семендяев К. А., О нахождении собственных значений и инвариантных многообразий матриц посредством итераций, Прикл. мат. и мех. 3, 1943, 193—221.
 - [8] Лопшиц А. М., Численный метод нахождения собственных значений и собственных плоскостей линейного оператора, Труды сем. по вект. и тензорн. исчисл., т. VII, 233—259.
 - [9] Канторович Л. В., Функциональный анализ и прикладная математика, Усп. матем. наук 3:6 (28) (1948), 89—185.
 - [10] Бирман М. Ш., Некоторые оценки для метода наискорейшего спуска, Усп. матем. наук 5, вып. 3 (40), 1950, 152—155.
 - [11] Гавурин М. К., Применение полиномов наилучшего приближения к улучшению сходимости итерационных процессов, Усп. матем. наук 5, вып. 3 (40), 1950, 156.
 - [12] Лопшиц А. М., Экстремальная теорема для гиперэллипсоида и ее применение к решению системы линейных уравнений, Труды сем. по вект. и тензорн. исчисл., т. IX, 1952, 183—197.
 - [12a] Красносельский М. А. и Крейн С. Г., Итеративный процесс с минимальными невязками, Матем. сб., 31 (173), 1952, 315—334.
 - [13] Шрейдер Ю. А., Решение систем линейных алгебраических уравнений, ДАН 76, 651—655.
 - [14] Lanczos C., Linear systems in self adjoint Form, Am. Math. M. 65, 1958, 665—679.
 - [15] Lancros C., Iterative solution of large scale leneare systems. Soc. Indust. Appl. Math., 6, 1, 1958, 91—109.
 - [16] Загускин В. Л., О решении систем линейных уравнений методами Л. В. Канторовича и А. М. Лопшица, Уч. зап. Яросл. пед. инст., в. XXXIV, ч. II, 67—80.
 - [17] Гершгорн С. А., Über die Abgrenzung der Eigenwerte einer Matrix, Изв. АН СССР, сер. матем., 7, 1931, 749—754.
-

ГЛАВА IV

ГАРМОНИЧЕСКИЙ АНАЛИЗ

1. Исторические замечания

Открытие акустического значения основного колебания и его обертонов приписывают греческому мудрецу Пифагору (600 лет до н. э.). Математическое значение разложения периодической функции в ряд по функциям вида $\sin kx$ и $\cos kx$, где k — целое число, было осознано в XVIII веке Эйлером и Лагранжем. Так как тогда математики не владели еще точным понятием предела, то истинная природа бесконечных рядов была им еще не вполне ясна. Лагранж считал, что всякая суперпозиция аналитических (т. е. бесконечно дифференцируемых) функций должна снова давать аналитическую функцию. Поэтому он полагал, что гармонический анализ можно применять лишь к функциям, которые не только непрерывны, но имеют, кроме того, производные любого порядка.

Блестящим открытием Фурье [«*Theorie analytique de la chaleur*» (1822)] было установление того факта, что это ограничение излишне. Функция может быть совершенно «неформульной», т. е. состоящей из произвольного числа «кусков», аналитическое выражение которых совершенно различно. Необязательна также непрерывность функций. Фурье дал примеры гармонического анализа функций, имевших конечное число разрывов в заданном основном интервале. Значения, принимаемые функцией в этом основном интервале, обычно нормируемом до интервала $(-\pi, +\pi)$ — это все, что необходимо для определения функции $y = f(x)$. Периодичность функции является несущественным свойством и влияет на ее изображение только при выходе за пределы основного интервала. Теория гармонического анализа не требует выхода за пределы основного интервала.

Другим фундаментальным открытием Фурье был «интеграл Фурье», с помощью которого методы гармонического анализа обобщались на случай бесконечного интервала с границами $(-\infty, +\infty)$, причем не требовалось какой-либо периодичности разлагаемой функции. Позднее Дирихле исследовал более детально, каким условиям должна удовлетворять «произвольная функция» Фурье, чтобы она могла быть представлена гармоническим рядом. Эти условия, так называемые

«условия Дирихле», достаточны для сходимости ряда Фурье; однако они не всегда необходимы. В новейшее время (1904) Фейер нашел новый метод суммирования рядов Фурье, который значительно расширяет область их применения. Пользуясь арифметическими средними частных сумм, вместо самих частных сумм, он сумел суммировать ряды, которые расходились сами по себе; единственное условие, которому функция должна удовлетворять,— это естественное условие, что она должна быть абсолютно интегрируемой.

2. Основные теоремы

Пусть функция $y=f(x)$ удовлетворяет условиям:

1. $f(x)$ определена в каждой точке интервала

$$-\pi \leq x \leq +\pi.$$

2. $f(x)$ всюду однозначна, конечна и кусочно непрерывна. Следовательно, $f(x)$ может иметь лишь конечное число разрывов, и два последовательных разрыва должны быть отделены конечным интервалом.

3. $f(x)$ имеет «ограниченную вариацию»; это значит, что $f(x)$ не может иметь бесконечного числа максимумов и минимумов в заданном интервале.

Эти условия, налагаемые на $f(x)$, называют «условиями Дирихле». Функция, удовлетворяющая условиям Дирихле, может быть разложена в сходящийся бесконечный ряд вида

$$f(x) = \frac{1}{2} a_0 + a_1 \cos x + a_2 \cos 2x + \dots + \\ + b_1 \sin x + b_2 \sin 2x + \dots, \quad (4-2.1)$$

где коэффициенты a_k , b_k , называемые «коэффициентами Фурье», определяются формулами

$$\left. \begin{aligned} a_k &= \frac{1}{\pi} \int_{-\pi}^{+\pi} f(x) \cos kx \, dx, \\ b_k &= \frac{1}{\pi} \int_{-\pi}^{+\pi} f(x) \sin kx \, dx. \end{aligned} \right\} \quad (4-2.2)$$

В точке разрыва $x=a$ мы определим $f(a)$ как среднее арифметическое двух предельных ординат

$$f(a) = \frac{1}{2} [f(a_+) + f(a_-)]^* \quad (4-2.3)$$

*) $f(a_+) = \lim_{\epsilon \rightarrow 0} f(a + \epsilon)$, $f(a_-) = \lim_{\epsilon \rightarrow 0} f(a - \epsilon)$, когда ϵ , оставаясь положительным, стремится к нулю.

Это то значение, к которому сходится соответствующий бесконечный ряд Фурье в точке $x=a$.

Удобно записывать ряд Фурье и в следующей комплексной форме:

$$f(x) = c_0 + c_1 e^{ix} + c_2 e^{2ix} + \dots + \\ + c_{-1} e^{-ix} + c_{-2} e^{-2ix} + \dots = \sum_{k=-\infty}^{+\infty} c_k e^{ikx}, \quad (4-2.4)$$

где

$$c_k = \frac{1}{2\pi} \int_{-\pi}^{+\pi} f(x) e^{-ikx} dx. \quad (4-2.5)$$

Доказательство. Если принять, что ряд (1) сходится к $f(x)$ в каждой точке заданного интервала, то можно умножить обе части равенства (1) на $\cos kx$ или $\sin kx$ и проинтегрировать от $-\pi$ до $+\pi$. Тогда мы тотчас же получим выражения (2) для коэффициентов разложения a_k, b_k в силу того, что в правой части при интегрировании все члены исчезают, за исключением одного, содержащего $\cos kx$ или $\sin kx$. Эти формулы (2) очень просты и были хорошо известны задолго до Фурье. Но вправе ли мы считать обоснованным разложение (1)? Чтобы ответить на этот вопрос, Дирихле рассуждал следующим образом. Рассмотрим *конечный* ряд

$$f_n(x) = \frac{1}{2} a_0 + a_1 \cos x + \dots + a_n \cos nx + \\ + b_1 \sin x + \dots + b_n \sin nx, \quad (4-2.6)$$

считая, что a_k и b_k определены формулами (2), и исследуем, что происходит, когда n бесконечно возрастает. Суммируя (6), получаем

$$f_n(x) = \int_{-\pi}^{+\pi} f(t) K_n(x-t) dt, \quad (4-2.7)$$

где $K_n(\xi)$ есть так называемое «ядро Дирихле»:

$$K_n(\xi) = \frac{\sin\left(n + \frac{1}{2}\right)\xi}{2\pi \sin\left(\frac{1}{2}\xi\right)}. \quad (4-2.8)$$

Наша цель состоит в том, чтобы показать, что с возрастанием n до бесконечности $f_n(x)$ неограниченно приближается к $f(x)$, т. е. что разность между этими величинами может быть сделана произвольно малой. Для такого поведения $f_n(x)$ необходимо требуется весьма сильное *фокусирующее действие* функции $K_n(\xi)$: она должна выделить непосредственную окрестность точки $\xi=0$ и стереть все

остальное. В частности, можно полагать, что следующие два свойства являются математическим выражением возрастающего фокусирующего действия функции $K_n(\xi) = K_n(-\xi)$:

$$\lim_{n \rightarrow \infty} \int_{\varepsilon}^{\pi} |K_n(\xi)| d\xi = 0, \quad (4-2.9)$$

$$\lim_{n \rightarrow \infty} \int_{-\varepsilon}^{+\varepsilon} |K_n(\xi)| d\xi = 1. \quad (4-2.10)$$

Первое условие обеспечивает, что функция $K_n(\xi)$ *стирает* в формуле (7) все, кроме ближайшей окрестности точки $t = x$. Второе условие обеспечивает, что точка $t = x$ участвует в интегрировании с надлежащим весом (предполагается, что ε есть заранее выбранное произвольно малое положительное число, не зависящее от n).

Однако ядро Дирихле не удовлетворяет первому условию. Вторичные максимумы функции (2.8) недостаточно малы, чтобы обеспечить преобладающую роль точки $\xi = 0$. Это обстоятельство является причиной того, что возможность разложения функции $f(x)$ в бесконечный сходящийся ряд Фурье должна быть ограничена определенным классом функций, которые достаточно гладки, чтобы противодействовать недостаточно фокусирующему действию ядра Дирихле. Условия Дирихле 2 и 3, приведенные выше, и устанавливают желаемую гладкость функции¹⁾.

Метод Фейера. Остроумным видоизменением процесса суммирования Фейеру удалось усилить фокусирующее действие ядра Дирихле и тем самым расширить область пригодности рядов Фурье на более обширный класс функций²⁾. Ограничим класс функции $f(x)$ лишь единственным условием, что $f(x)$ «абсолютно интегрируема»; это условие означает существование интеграла

$$I_1 = \int_{-\pi}^{+\pi} |f(x)| dx. \quad (4-2.11)$$

Такое ограничение совершенно естественно, ибо иначе коэффициенты Фурье a_k и b_k , определенные формулами (2), или c_k , определенные формулой (5), не существовали бы.

Теперь можно формально построить ряды (1) или (4). Вообще говоря, эти ряды не будут сходиться. Это значит, что последовательность функций $f_n(x)$, определенных равенством (6), не будет

¹⁾ Условия Дирихле достаточны; однако они налагают на функцию излишне жесткие ограничения, хотя и включают весьма широкий класс функций. Минимум ограничений, необходимых для сходимости ряда Фурье, еще неизвестен. В методе Фейера этот вопрос обходится, ибо иначе суммируется ряд.

²⁾ Ср. {15}, стр. 169.

стремиться к определенному пределу при бесконечном возрастании числа n . Функции $f_n(x)$ называют «частными суммами», так как они образуются суммированием конечного числа членов ряда. Введем для $f_n(x)$ другое обозначение $s_n(x)$. Подставим в $s_n(x)$ определенное значение x и рассмотрим последовательность

$$s_0, s_1, s_2, s_3, \dots$$

Вычисляя средние арифметические членов последовательности, построим новую последовательность:

$$f_1 = s_0, \quad f_2 = \frac{s_0 + s_1}{2}, \quad \dots, \quad f_n = \frac{s_0 + s_1 + \dots + s_{n-1}}{n}. \quad (4-2.12)$$

Эта новая последовательность обладает замечательным свойством: она сходится к определенному пределу во всех точках, в которых существуют односторонние пределы $f_1 = f(x_+)$ и $f_2 = f(x_-)$, определяемые формулами

$$\left. \begin{aligned} \lim_{\varepsilon \rightarrow 0} [f(x + \varepsilon) - f_1] &= 0, \\ \lim_{\varepsilon \rightarrow 0} [f(x - \varepsilon) - f_2] &= 0 \end{aligned} \right\} \quad (\varepsilon > 0). \quad (4-2.13)$$

Во всех таких точках $f_n(x)$ стремится к арифметическому среднему двух пределов (13) [в обыкновенной точке эти два предела совпадают, и ряд просто сходится к $f(x)$]. Следовательно, $f_n(x)$ стремится к разумному значению во всех точках, в которых можно вообще ожидать разумного ответа.

Результаты Фейера вытекают из того факта, что его метод приводит к следующей функции, представляющей ядро

$$K_n(\xi) = \frac{\sin^2\left(\frac{n\xi}{2}\right)}{2\pi n \sin^2\left(\frac{\xi}{2}\right)}. \quad (4-2-14)$$

Это ядро обладает сильными фокусирующими свойствами, выраженными формулами (9) и (10).

3. Квадратичные приближения

В то время как интересы чистого анализа существенно сосредоточены на общих свойствах сходимости ряда Фурье, интересы практического анализа сосредоточены на возможности хорошего представления достаточно обширного, но не слишком иррегулярного класса функций с помощью относительно небольшого числа гармонических составляющих. Здесь уже вопрос идет не о представлении заданной функции $y = f(x)$ с произвольно малой ошибкой, а об аппроксими-

ровании $f(x)$ с конечной ошибкой, которую, однако, надо сделать возможно малой.

Ряд Фурье принадлежит к тому классу разложений, который аппроксимирует заданную функцию $y = f(x)$ в определенном конечном интервале «равномерно хорошо» — это означает, что мы не стараемся получить хорошее приближение в малой окрестности некоторой точки за счет того, что вне этой окрестности получается большое расхождение, а относимся с одинаковым вниманием ко всем точкам интервала.

Руководящий принцип такого рода интерполяции дается «методом наименьших квадратов». Мы определяем ошибку $\eta(x)$ разложения в ряд с конечным числом членов, образуя разность между заданной функцией $y = f(x)$ и конечной суммой

$$f_m(x) = c_1\varphi_1(x) + c_2\varphi_2(x) + \dots + c_m\varphi_m(x), \quad (4-3.1)$$

где $\varphi_1(x)$, $\varphi_2(x)$, ..., $\varphi_m(x)$ суть заранее выбранные функции, а коэффициентами мы можем распорядиться. Эту разность обозначаем через $\eta_m(x)$:

$$\eta_m(x) = f(x) - f_m(x). \quad (4-3.2)$$

Вместо того чтобы стремиться сделать $\eta_m(x)$ равным нулю в m более или менее произвольно выбранных точках *), мы охарактеризуем «среднюю ошибку» путем интегрирования в конечном интервале

$$a \leq x \leq b, \quad (4-3.3)$$

в котором функция $f(x)$ должна быть представлена. Мы определяем эту среднюю ошибку определенным интегралом

$$\bar{\eta}^2 = \frac{1}{b-a} \int_a^b \eta_m^2(x) dx \quad (4-3.4)$$

и находим коэффициенты c_i разложения (1) из условия, что ошибка $\bar{\eta}$ должна быть наименьшей из возможных. Эта задача всегда имеет определенное решение, так как существенно положительная функция всегда достигает определенного абсолютного минимума при некоторых значениях переменных.

Наша задача весьма облегчается, если заданные базисные функции

$$\varphi_1, \varphi_2, \dots, \varphi_n \quad (4-3.5)$$

удовлетворяют следующему «условию ортогональности»:

$$\int_a^b \varphi_i(x) \varphi_k(x) dx = 0 \quad (i \neq k). \quad (4-3.6)$$

*) Как обычно поступают при интерполяции.

Семейство функций, удовлетворяющих этому условию, называется «ортогональным семейством». Если, кроме того, мы выберем для каждой функции $\varphi_k(x)$ такой нормирующий множитель, чтобы имело место равенство

$$\int_a^b \varphi_i^2(x) dx = 1, \quad (4-3.7)$$

то получим «ортогональное и нормированное» или кратко «ортонормированное» семейство функций.

Заметим, что для такой системы функций интеграл от $f_m^2(x)$, взятый в данном интервале, принимает особенно простой вид

$$\int_a^b f_m^2(x) dx = c_1^2 + c_2^2 + \dots + c_m^2. \quad (4-3.8)$$

Далее интеграл от $\eta_m^2(x)$ также записывается очень просто:

$$\begin{aligned} \int_a^b \eta_m^2(x) dx = \int_a^b f^2(x) dx - 2(\gamma_1 c_1 + \gamma_2 c_2 + \dots + \gamma_m c_m) + \\ + c_1^2 + c_2^2 + \dots + c_m^2, \end{aligned} \quad (4-3.9)$$

где

$$\gamma_k = \int_a^b f(x) \varphi_k(x) dx. \quad (4-3.10)$$

Но выражение (9) есть простая алгебраическая функция переменных c_i , и легко убедиться, что она принимает минимальное значение, если

$$c_i = \gamma_i. \quad (4-3.11)$$

При этом выборе c_i средняя ошибка $\bar{\eta}_m$ приближения определится по формуле

$$\bar{\eta}_m^2 = \frac{1}{b-a} \left[\int_a^b f^2(x) dx - (\gamma_1^2 + \gamma_2^2 + \dots + \gamma_m^2) \right]. \quad (4-3.12)$$

Если заданная функция $f(x)$ есть как раз линейная комбинация m выбранных функций $\varphi_k(x)$, то выбор коэффициентов c_i по формуле (11) дает разложение (1), которое тождественно с $f(x)$. В этом случае $\bar{\eta}_m$ окажется равным нулю. Однако при всяком другом выборе функции $f(x)$ $\bar{\eta}_m^2$ останется положительной величиной, которая может быть дополнительно уменьшена только путем добавления новых функций $\varphi_{m+1}(x)$, ... к прежней системе.

Большое преимущество ортогональной системы функций состоит в том, что добавление нового члена к системе не оказывает вредного влияния на прежние вычисления. Нет необходимости что-либо менять в прежних коэффициентах, следует только определить дополнительный коэффициент c_{m+1} , который зависит лишь от $\varphi_{m+1}(x)$.

Средняя ошибка η_{m+1} теперь становится еще меньше, так как к слагаемым, заключенным внутри скобок правой части (12), добавляется $-\gamma_{m+1}^2$. Если γ_{m+1} оказывается равным нулю, то добавление функции $\varphi_{m+1}(x)$ не уменьшает среднюю ошибку; тогда нужно попробовать $\varphi_{m+2}(x)$ и т. д.

4. Ортогональность функций Фурье

Функции ряда Фурье представляют собой пример ортогонального семейства. Интервал здесь берется от $-\pi$ до $+\pi$ или от 0 до 2π , и мы имеем

$$\int_{-\pi}^{\pi} e^{imx} e^{inx} dx = \frac{1}{i(m+n)} [e^{i(m+n)x}]_{-\pi}^{+\pi} = 0 \quad (m+n \neq 0). \quad (4-4.1)$$

Однако для функций, принимающих комплексные значения, понятие ортогональности следует слегка видоизменить. Здесь величина $\eta_m^2(x)$ заменяется произведением $\eta_m(x) \eta_m^*(x)$, где знак * обозначает переход от функции $\eta_m(x) = p(x) + iq(x)$ (где $p(x)$, $q(x)$ — вещественные числа) к функции $\eta_m^*(x) = p(x) - iq(x)$. Соответственно этому ортогональность означает теперь, что

$$\int_a^b \varphi_i(x) \varphi_k^*(x) dx = 0 \quad (i \neq k), \quad (4-4.2)$$

а нормирование функций $\varphi_i(x)$ производится с помощью условия

$$\int_a^b \varphi_i(x) \varphi_i^*(x) dx = 1. \quad (4-4.3)$$

Далее, коэффициенты γ_k теперь определяются формулой

$$\gamma_k = \int_a^b f(x) \varphi_k^*(x) dx, \quad (4-4.4)$$

и эти величины являются коэффициентами ортогонального разложения*), которые минимизируют среднюю ошибку.

*) То есть разложения по ортогональным функциям.

Ортонормированными функциями комплексного ряда Фурье являются функции

$$\varphi_k(x) = \frac{1}{\sqrt{2\pi}} e^{ikx}, \quad (4-4.5)$$

а для коэффициентов γ_k получаем выражение

$$\gamma_k = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{+\pi} f(x) e^{-ikx} dx. \quad (4-4.6)$$

Результирующий ряд совпадает с указанной выше комплексной формой (2.4), (2.5) ряда Фурье.

Заметим, что разложение в бесконечный ряд Фурье (2.4) можно понимать как процесс последовательного уменьшения средней ошибки путем добавления все новых и новых ортогональных функций. Можно показать, что с возрастанием n до бесконечности средняя ошибка [ср. (3.4)] стремится к нулю для всякой «квадратичной интегрируемой функции» в этом интервале, т. е. для всякой функции, для которой существует интеграл

$$I_2 = \int_{-\pi}^{+\pi} |f(x)|^2 dx. \quad (4-4.7)$$

И все же имеет место то своеобразное явление, что из этого факта мы не можем вывести каких-либо заключений относительно равенства нулю самой локальной ошибки $\eta(x)$. Если бы средняя ошибка полностью исчезла, это необходимо означало бы исчезновение функции $\eta(x)$ в каждой точке, так как интеграл от всюду неотрицательной функции может равняться нулю только тогда, когда функция тождественно равна нулю. Такой факт обозначал бы, что ряд Фурье дает точное значение функции в каждой точке интервала. Однако мы имеем возможность только доказать, что средняя ошибка при достаточно большом k может быть сделана *сколь угодно малой*. Этого недостаточно для того, чтобы сама локальная ошибка стремилась к нулю в каждой точке интервала. Для выполнения такой локальной сходимости недостаточно, чтобы функция $f(x)$ была квадратично интегрируема (необходимо усилить это требование, например потребовав выполнения условий Дирихле, см. § 2).

5. Отделение ряда синусов от ряда косинусов

Для многих задач прикладного анализа целесообразно разделить часть ряда Фурье с синусами от части с косинусами. Этого можно достигнуть разбиением функции $f(x)$ на четную и нечетную части.

Положим

$$\left. \begin{aligned} g(x) &= \frac{1}{2} [f(x) + f(-x)], \\ h(x) &= \frac{1}{2} [f(x) - f(-x)]. \end{aligned} \right\} \quad (4-5.1)$$

Тогда

$$f(x) = g(x) + h(x), \quad (4-5.2)$$

где в силу (1)

$$g(-x) = g(x), \quad (4-5.3)$$

$$h(-x) = -h(x). \quad (4-5.4)$$

Функция со свойством симметрии (3) называется «четной», а со свойством (4) — «нечетной».

Для четной функции из разложения в ряд Фурье выпадает та часть, которая содержит синусы, для нечетной функции выпадает часть, содержащая косинусы. В обоих случаях достаточно задать функцию только в интервале

$$0 \leq x \leq \pi, \quad (4-5.5)$$

представляющем половину интервала полного ряда Фурье. В другой половине интервала значение функции определяется уже из свойств отражения (3) и (4).

Таким образом, мы получаем два разложения:

$$g(x) = \frac{1}{2} a_0 + a_1 \cos x + a_2 \cos 2x + \dots, \quad (4-5.6)$$

где

$$a_k = \frac{2}{\pi} \int_0^{\pi} g(x) \cos kx \, dx, \quad (4-5.7)$$

и

$$h(x) = b_1 \sin x + b_2 \sin 2x + \dots, \quad (4-5.8)$$

где

$$b_k = \frac{2}{\pi} \int_0^{\pi} h(x) \sin kx \, dx. \quad (4-5.9)$$

С другой стороны, можно исходить из предположения, что функция $f(x)$ с самого начала определена лишь в интервале (5). В этом случае у нас есть свободная возможность определить функцию для отрицательных значений x либо условием (3), либо условием (4). Следовательно, одна и та же функция $f(x)$ в интервале $(0, \pi)$ может быть разложена либо в ряд косинусов, либо в ряд синусов. Оба разложения сходятся в интервале $(0, \pi)$ к заданной функции $f(x)$, если только $f(x)$ принадлежит к классу функций, охватываемых условиями Дирихле.

Очень часто приходится иметь дело с разложением в ряд Фурье функции, которая в интервале $(0, \pi)$ всюду непрерывна и даже дифференцируема. В этом случае характер сходимости ряда Фурье определится свойствами функции в граничных точках $x=0$ и $x=\pi$. Если значения функции $f(x)$ в этих двух граничных точках не равны нулю, то отражение по принципу нечетной функции (4) приведет к разрыву в двух точках: $x=0$ и $x=\pi$. Разрыв в точке $x=0$ вытекает непосредственно из закона отражения (4), а разрыв в точке $x=\pi$ следует из условия периодичности

$$f(x + 2\pi) = f(x), \quad (4-5.10)$$

которое в применении к точке $x = -\pi$ требует, чтобы

$$f(\pi) = f(-\pi). \quad (4-5.11)$$

Оба разрыва можно ликвидировать, если отразить $f(x)$ как *четную* функцию. Тогда разрыв появляется лишь в *первой производной*, но не в самой функции. По этой причине для функции, не связанной особыми условиями на границе, разложение в ряд (6) по косинусам будет обладать гораздо лучшими свойствами сходимости, чем разложение (8) по синусам. В последнем разложении будут сказываться нежелательные колебания, вызванные «явлением Гиббса», которое возникает в окрестности точки разрыва (см. § 9).

С другой стороны, допустим, что $f(x)$ в обеих конечных точках интервала $(0, \pi)$ равна нулю. Этого всегда можно достигнуть, если вычесть из $f(x)$ линейную функцию $\alpha + \beta x$. При надлежащем подборе α и β новая функция будет удовлетворять краевым условиям

$$f(0) = 0, \quad f(\pi) = 0. \quad (4-5.12)$$

Для такого рода функции разложение в ряд по синусам дает гораздо лучшую сходимость, чем разложение в ряд косинусов, так как отражение типа нечетной функции обеспечивает непрерывность и самой функции, и *первой ее производной*; разрывы в точках $x=0$ и $x=\pi$ появляются лишь во *второй производной*.

Оценим теперь порядок величин коэффициентов ряда Фурье. При малом k о величине интеграла

$$\int f(x) \cos kx \, dx \quad (4-5.13)$$

немного можно сказать. Однако при больших значениях k первый множитель достаточно гладок по сравнению со вторым, который очень быстро колеблется.

Проинтегрировав по частям, получим

$$\int f(x) \cos kx \, dx = \frac{f(x) \sin kx}{k} - \frac{1}{k} \int f'(x) \sin kx \, dx. \quad (4-5.14)$$

Повторив процесс еще раз, найдем

$$\int f'(x) \sin kx \, dx = -\frac{f'(x) \cos kx}{k} + \frac{1}{k} \int f''(x) \cos kx \, dx. \quad (4-5.15)$$

Вычисляя интеграл в пределах от 0 до π , получим для доминирующего члена при больших значениях k приближенное равенство

$$\int_0^\pi f(x) \cos kx \, dx = \frac{1}{k^2} \left| f'(x) \cos kx \right|_0^\pi = \frac{(-1)^k f'(\pi) - f'(0)}{k^2}. \quad (4-5.16)$$

Аналогичная выкладка дает

$$\int_0^\pi f(x) \sin kx \, dx = -\frac{1}{k} \left| f(x) \cos kx \right|_0^\pi = \frac{f(0) - (-1)^k f(\pi)}{k}. \quad (4-5.17)$$

Из этих формул видно, что коэффициенты ряда косинусов убывают со скоростью k^{-2} , а коэффициенты ряда синусов — только со скоростью k^{-1} , если $f(x)$ не удовлетворяет специальным граничным условиям.

Если, однако, $f(x)$ удовлетворяет граничным условиям (12), то правая часть равенства (17) равна нулю, и для доминирующего члена получаем

$$\int_0^\pi f(x) \sin kx \, dx = \frac{1}{k^3} \left| f''(x) \cos kx \right|_0^\pi = \frac{(-1)^k f''(\pi) - f''(0)}{k^3}. \quad (4-5.18)$$

Здесь коэффициенты убывают со скоростью k^{-3} , которая является удовлетворительной для многих приложений рядов Фурье.

Мы можем перенести наши результаты, полученные для интервала $[0, \pi]$, на случай интервала $[-\pi, \pi]$. Для этого мы обозначим прежнюю переменную через x_1 и перейдем к новой переменной путем преобразования

$$x = -\pi + 2x_1. \quad (4-5.19)$$

Мы получаем таким образом для функции $f(x)$, удовлетворяющей граничным условиям

$$f(\pm \pi) = 0, \quad (4-5.20)$$

разложение

$$\left. \begin{aligned} f(x) = & a_1 \cos \frac{x}{2} + a_2 \cos \frac{3x}{2} + \dots + a_k \cos \frac{2k-1}{2} x + \\ & + \dots + b_1 \sin x + b_2 \sin 2x + \dots + b_k \sin kx + \dots, \end{aligned} \right\} \quad (4-5.21)$$

где

$$\left. \begin{aligned} a_k &= \frac{1}{\pi} \int_{-\pi}^{+\pi} f(x) \cos \frac{2k-1}{2} x dx, \\ b_k &= \frac{1}{\pi} \int_{-\pi}^{+\pi} f(x) \sin kx dx. \end{aligned} \right\} \quad (4-5.22)$$

Ряд (21) сходится быстрее, чем первоначальный ряд Фурье (2.1), если функция удовлетворяет граничным условиям (20); его коэффициенты убывают со скоростью k^{-3} , в то время как при разложении в первоначальный ряд Фурье (2.1) они убывали со скоростью k^{-2} .

6. Дифференцирование ряда Фурье

Рассмотрим сумму конечного числа членов ряда Фурье — частную сумму

$$f_m(x) = \sum_{k=-(m-1)}^{k=m-1} c_k e^{ikx} \quad (4-6.1)$$

вместе с остатком

$$\eta_m(x) = \sum_{k=m}^{\infty} (c_k e^{ikx} + c_{-k} e^{-ikx}). \quad (4-6.2)$$

Этот остаток можно записать в виде

$$\eta_m(x) = e^{imx} \sum_{k=0}^{\infty} c_{m+k} e^{ikx} + e^{-imx} \sum_{k=0}^{\infty} c_{-m-k} e^{-ikx}. \quad (4-6.3)$$

Положим

$$\sum_{k=0}^{\infty} c_{m+k} e^{ikx} = \rho_m(x). \quad (4-6.4)$$

Функция $\rho_m(x)$ есть, вообще говоря, гладкая функция, не обнаруживающая сколько-нибудь быстрых колебаний. С другой стороны, e^{imx} есть быстро колеблющаяся функция. Поэтому остаток ряда Фурье имеет характер «модулированной несущей волны» высокой частоты.

Если мы формально продифференцируем усеченный ряд (1) и сравним эту производную с $f'(x)$, то получим большую погрешность

$$\eta'_m(x) = im e^{imx} \rho_m(x) + e^{imx} \rho'_m(x) - im e^{-imx} \rho_{-m}(x) + e^{-imx} \rho'_{-m}(x). \quad (4-6.5)$$

Штрихованные члены в этом выражении получаются от дифференцирования модуляции $\rho_m(x)$ и не вызывают серьезных затруднений. Однако другие члены содержат большой множитель m , и именно это

обстоятельство обуславливает серьезную потерю в точности при формальном дифференцировании ряда Фурье.

Мы дифференцировали и модуляцию, и несущую волну. Нам надо было бы дифференцировать только модуляцию без несущей волны. Этого мы можем достичь с помощью следующего приема. Заменяем процесс дифференцирования следующим разностным процессом:

$$\mathcal{D}_m y = \frac{f\left(x + \frac{\pi}{m}\right) - f\left(x - \frac{\pi}{m}\right)}{\frac{2\pi}{m}}. \quad (4-6.6)$$

Оператор $\mathcal{D}_m$ не совпадает с оператором D обычного дифференцирования, но различие становится все меньше и меньше по мере возрастания m до бесконечности. В пределе

$$D = \frac{d}{dx} = \lim_{m \rightarrow \infty} \mathcal{D}_m. \quad (4-6.7)$$

Но оператор $\mathcal{D}_m$, примененный к $\eta_m(x)$, отбирает две точки на несущей волне, находящиеся точно в одинаковой фазе, а именно, отличающиеся на $\pm 180^\circ$ от фазы точки x . Ввиду этого оператор $\mathcal{D}_m$ добавляет лишь множитель -1 к несущей волне, не дифференцируя ее. Мы получаем

$$\mathcal{D}_m \eta_m(x) = -e^{imx} \mathcal{D}_m \rho_m(x) - e^{-imx} \mathcal{D}_m \rho_{-m}(x). \quad (4-6.8)$$

Но $\rho_m(x)$ и $\rho_{-m}(x)$ суть гладкие функции, дифференцирование которых не вызывает увеличения порядка величины погрешности. Следовательно, сходимость ряда Фурье не ухудшилась от этой операции. В пределе, когда m возрастает до бесконечности, мы получаем производную от $f(x)$ во всех точках, где она существует.

Этот метод дифференцирования ряда Фурье эквивалентен умножению коэффициентов формально дифференцированного ряда на некоторые заранее установленные весовые множители σ_k , которые зависят от числа m порядка частных сумм. Рассмотрим член, соответствующий e^{ikx} :

$$\mathcal{D}_m e^{ikx} = i \frac{\sin \frac{\pi k}{m}}{\frac{\pi}{m}} e^{ikx} = \frac{\sin \frac{\pi k}{m}}{\frac{\pi k}{m}} i k e^{ikx}. \quad (4-6.9)$$

Поправочный множитель, таким образом, равен

$$\sigma_k = \frac{\sin \frac{\pi k}{m}}{\frac{\pi k}{m}}. \quad (4-6.10)$$

Так как то же значение σ_k получается и при $-k$, то один и тот же множитель получают как коэффициент c_k , так и c_{-k} . Но коэффициенты a_k , b_k вещественной формы ряда Фурье являются линейной комбинацией коэффициентов c_k и c_{-k} :

$$a_k = c_k + c_{-k}, \quad b_k = i(c_k - c_{-k}). \quad (4-6.11)$$

Поэтому коэффициенты a_k и b_k получают подобным же образом одинаковые «множители затухания» σ_k . Множитель σ_k равен 1 лишь при $k=0$; с увеличением k величина σ_k уменьшается монотонно и становится почти равной нулю при $k=m-1$.

Сильное затухание коэффициентов Фурье высокого порядка успешно противодействует их тенденции делать ряд расходящимся. Поэтому ряд, который был бы при обычном дифференцировании совершенно расходящимся, может быть сделан сходящимся путем применения σ -множителей.

7. Разложение дельта-функции в тригонометрический ряд

Хорошим примером может служить «квадратная волна», определенная следующими условиями:

$$\left. \begin{aligned} f(-x) &= -f(x), \\ f(x) &= \frac{1}{2} \quad (0 < x < \pi), \\ f(0) &= f(\pi) = 0. \end{aligned} \right\} \quad (4-7.1)$$

Соответствующий ряд Фурье

$$y(x) = \frac{2}{\pi} \left(\sin x + \frac{\sin 3x}{3} + \frac{\sin 5x}{5} + \dots \right) \quad (4-7.2)$$

хотя и сходится в каждой точке интервала $(-\pi, +\pi)$ к $f(x)$, но сходимость эта очень медленная.

Если продифференцировать заданную функцию $f(x)$, то получим

$$f'(x) = 0 \quad (4-7.3)$$

в каждой точке, за исключением точки $x=0$, где производной не существует. С другой стороны, формальное дифференцирование ряда Фурье дает

$$y'(x) = \frac{2}{\pi} (\cos x + \cos 3x + \cos 5x + \dots). \quad (4-7.4)$$

Этот ряд расходится во всех точках интервала $(-\pi, +\pi)$, за исключением точек $x = \pm \frac{\pi}{2}$. Следовательно, сильная особенность в точке $x=0$ распространяет свое влияние на каждую точку области и разрушает сходимость ряда Фурье (4) всюду, хотя $f'(x)$ — вполне регулярная функция в каждой точке, исключая начало $x=0$.

Если теперь заменим ряд $y'(x)$ рядом

$$y'_{2m}(x) = \frac{2}{\pi} \left[\frac{\sin \frac{\pi}{2m}}{\frac{\pi}{2m}} \cos x + \frac{\sin \frac{3\pi}{2m}}{\frac{3\pi}{2m}} \cos 3x + \dots + \frac{\sin (2m-1) \frac{\pi}{2m}}{(2m-1) \frac{\pi}{2m}} \cos (2m-1)x \right], \quad (4-7.5)$$

то получим совсем другую картину. Когда m возрастает до бесконечности, $y'_{2m}(x)$ стремится к нулю в каждой фиксированной точке x , за исключением $x=0$, где функция возрастает беспредельно. При этом функция возрастает таким образом, что площадь, заключенная под той частью кривой, которая расположена между ординатами $\pm \frac{\varepsilon}{2}$, где ε — произвольно малая, но определенная величина, — стремится к 1.

Производную от функции (1) часто называют «дельта-функцией» и обозначают через $\delta(x)$. Это не обычная функция, ибо первоначальная функция не имеет производной в точке $x=0$. Она, однако, является пределом обычной функции и полезна для многих аналитических целей. Мы будем считать, что эта функция равна нулю повсюду, за исключением промежутка $\left(-\frac{\varepsilon}{2}, +\frac{\varepsilon}{2}\right)$, где ε стремится к нулю. В точке $x=0$ эта функция обращается в бесконечность таким образом, что площадь, заключенная под соответствующей частью кривой, равна 1.

Физической интерпретацией дельта-функции является импульс с единичной интенсивностью, действующий в течение промежутка времени ε близ точки $x=0$ при ε , стремящемся к нулю. Формальное определение дельта-функции приводит к бесконечному ряду

$$y = \frac{1}{\pi} \left(\frac{1}{2} + \cos x + \cos 2x + \dots \right), \quad (4-7.6)$$

который ни в какой точке не сходится. Если, однако, мы введем σ -множители и рассмотрим разложение

$$y_m = \frac{1}{\pi} \left(\frac{1}{2} + \sigma_1 \cos x + \sigma_2 \cos 2x + \dots + \sigma_{m-1} \cos (m-1)x \right), \quad (4-7.7)$$

где σ_k суть функции, определенные формулой (6.10), то получим ряд, который можно рассматривать как тригонометрическое представление дельта-функции. Рис. 24 дает наглядное изображение поведения полученной таким образом функции. Переменная ξ представляет здесь произведение

$$\xi = \frac{m}{\pi} x. \quad (4-7.8)$$

Следовательно, точки ± 1 на рис. 24 соответствуют значениям $\pm \frac{\pi}{m}$ переменной x . Кроме того, значения функции $y(\xi)$ находятся в следующем соотношении со значениями функции $y_m(x)$:

$$y_m(x) = my(\xi) = my\left(\frac{m}{\pi}x\right). \quad (4-7.9)$$

Множитель m стягивает кривую в направлении x и растягивает ее в направлении y . Когда m возрастает до бесконечности, y_m стремится

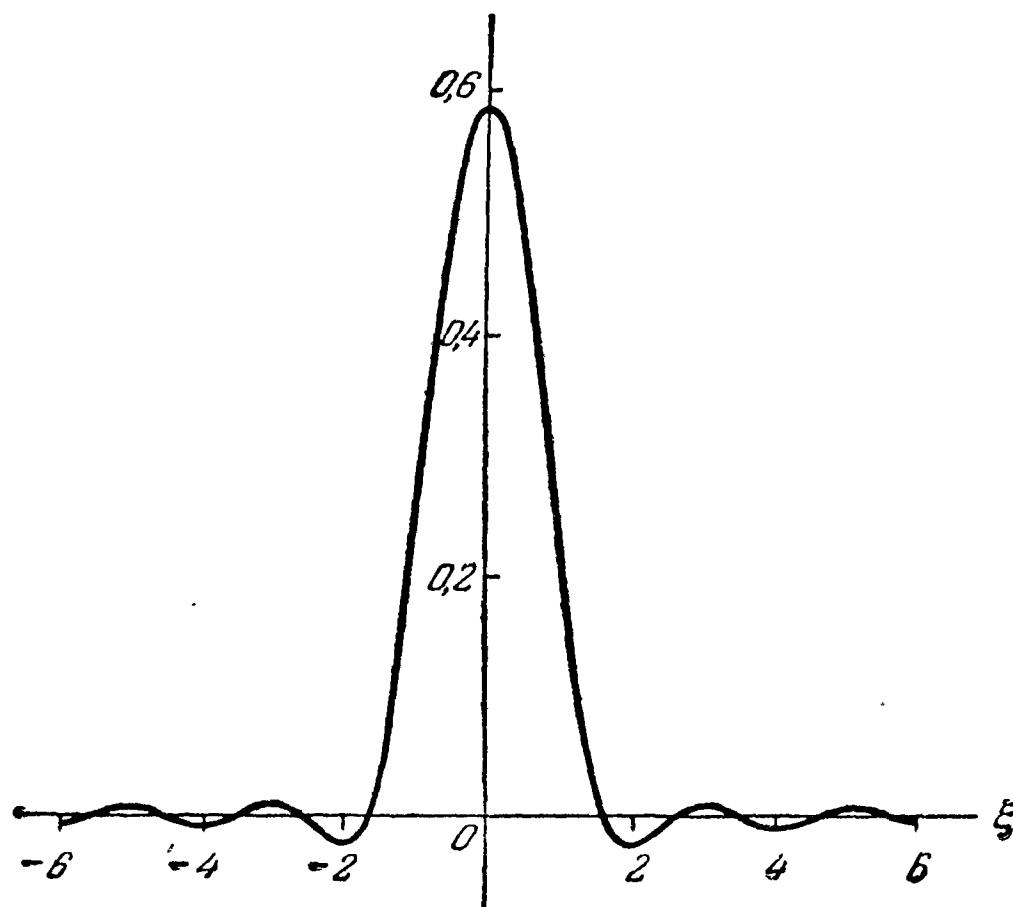


Рис. 24.

в каждой точке x к 0, за исключением точки $x=0$. В то же время площадь, расположенная под кривой, вычисленная с помощью ряда, становится равной 1.

8. Распространение тригонометрического ряда на неинтегрируемые функции

Как мы видели, для разложения функции $f(x)$ в ряд Фурье необходима абсолютная интегрируемость $f(x)$. Функцию $y = \log x$ можно разложить в интервале $|0, \pi|$ в ряд косинусов (или синусов), так как интеграл от этой функции ограничен, хотя сама функция обращается в бесконечность при $x=0$. С другой стороны, функция $y = x^{-1}$ не может быть разложена в ряд Фурье в этом интервале, так как она столь быстро возрастает, что даже площадь, заключенная под кривой, обращается в бесконечность. Коэффициенты Фурье a_k, b_k для такой функции становятся бесконечно большими.

Однако x^{-1} можно рассматривать как производную от функции $\log x$. Поскольку у нас имеется способ, который позволяет диффе-

ренцировать сходящийся ряд Фурье так, чтобы полученный ряд оставался сходящимся, то мы можем сначала разложить $y = \log x$ в ряд Фурье, а затем, продифференцировав его, получить ряд для x^{-1} . Это мы можем сделать, пользуясь множителями σ_k . Формальная производная ряда для $\log x$ нигде не сходится. Но если умножить формально полученные коэффициенты на $\sigma_k(m)$, то с возрастанием m измененный ряд сходится к функции x^{-1} , и погрешность может быть сделана произвольно малой во всех точках, кроме точки $x=0$, где производная не существует и функция обращается в бесконечность.

Ряд косинусов в этой точке точно так же обращается в бесконечность, тогда как ряд синусов дает нуль. Все же мы можем выбрать произвольно малое ϵ , и ряд синусов будет сходиться к весьма большому значению $\frac{1}{\epsilon}$ вопреки тому факту, что он должен начаться со значения нуль в точке $x=0$.

Многократное применение того же процесса дает простой метод получения тригонометрических разложений для функций, которые сами не интегрируемы, но представляют собой производные N -го порядка от абсолютно интегрируемых функций. В этом случае мы начинаем с этой образующей функции, получаем для нее ряд Фурье, формально дифференцируем его μ раз и затем умножаем коэффициенты этой производной функции на соответственным образом полученные множители σ_k . Измененный таким образом ряд будет сходиться к μ -й производной первоначальной функции в каждой точке, где эта производная существует. Пользуясь этим методом, можно разложить в тригонометрический ряд даже функцию типа $x^{-\mu}$ при произвольно большом положительном μ и погрешность сделать произвольно малой во всякой точке, кроме начала координат, хотя классического ряда Фурье для этих функций не существует.

9. Сглаживание колебаний Гиббса с помощью σ -множителей

Применение σ -множителей может служить не только для преобразования расходящегося ряда Фурье в сходящийся ряд, но также для улучшения сходимости слабо сходящегося ряда Фурье. Так как каждую функцию можно рассматривать как производную от своего интеграла, то уменьшение ошибки в производной благодаря множителю σ должно оказать благоприятное влияние на сходимость любого заданного ряда Фурье.

Медленная сходимость ряда Фурье особенно нежелательна, если функция имеет еще и точку разрыва. Влияние разрывности состоит в том, что возникают колебания вокруг истинного значения, и амплитуды этого колебания убывают очень медленно с увеличением числа членов. Колебания высокой частоты усеченного ряда вокруг истинного

значения функции всегда имеют место — это показывает выражение (6.3) для погрешности. Обыкновенно амплитуды этих колебаний достаточно малы, чтобы не вызвать особых последствий. Однако в случае разрыва они весьма заметны и мешают эффективному гармоническому синтезу «квадратной волны»¹⁾.

Метод арифметической средней Фейера полностью устраняет колебания Гиббса¹⁾. Приближение к квадратной волне, полученное по этому методу, осуществляется с помощью совершенно гладких монотонных функций, которые остаются постоянно ниже первоначальной кривой. Метод σ -множителей не избавляет от колебаний, но снижает их амплитуды. Первоначальная амплитуда выброса в первом колебании, составляющая 0,08949, уменьшается до 0,01187; ближайший минимум, равный 0,04859, доводится до 0,00473 и т. д.

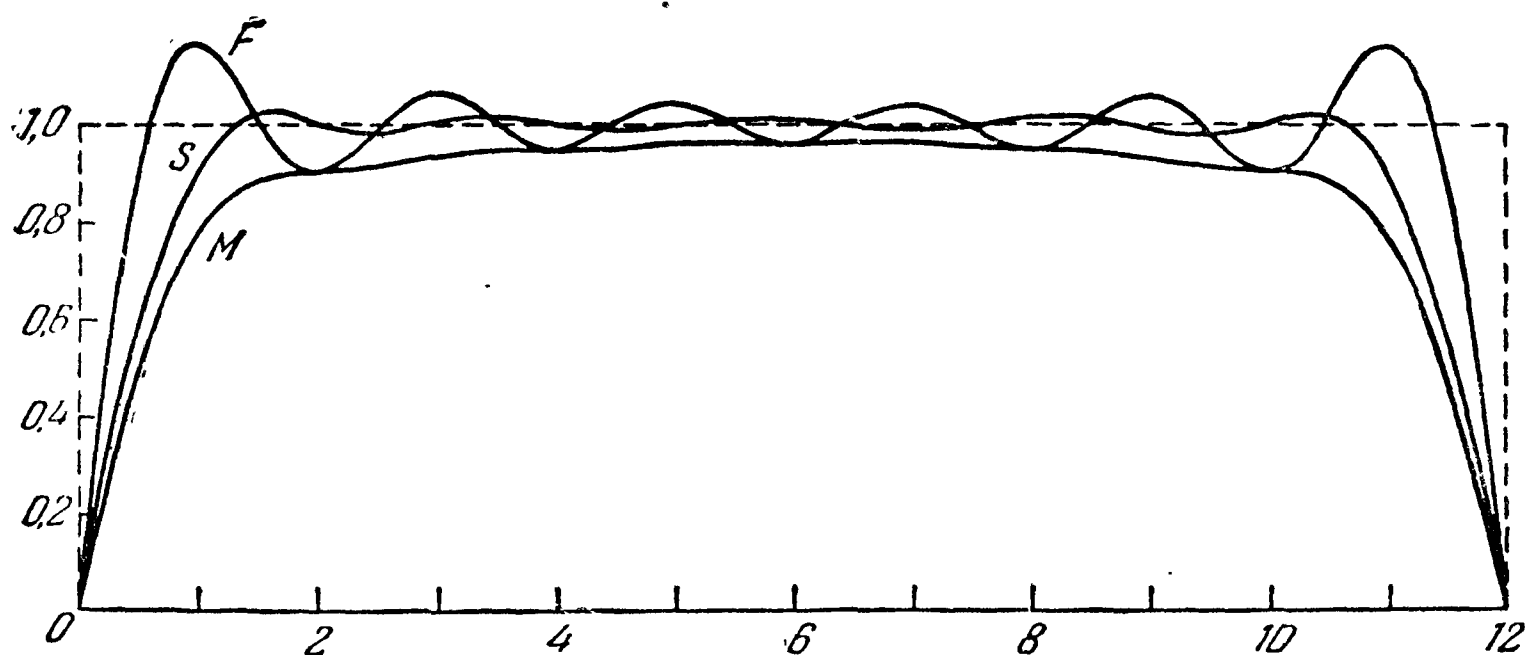


Рис. 25.

Прилагаемый рис. 25 изображает кривую F первоначальной усеченной суммы Фурье с колебаниями Гиббса, а также кривые арифметической средней M и результата сглаживания S с помощью σ -множителей. Отметим, что метод арифметической средней, устраняя нежелательные колебания ряда Фурье, обладает, однако, тем недостатком, что он дает весьма медленное приближение к асимптотическому значению, и соответствующая кривая начинается относительно малым подъемом. Метод σ -множителей даст более быстрое нарастание аппроксимирующей функции вместе с резким поворотом после того, как достигнут максимальный уровень. Правда, колебания все же остаются, но их амплитуда сильно заглушена и быстро затухает. Точность приближения в этом случае заметно лучше, чем точность при методе арифметической средней, хотя по сравнению с первоначальным рядом при последнем методе наблюдается практически полное устранение искажающего влияния явления Гиббса ценою несколько менее быстрого подъема в начале кривой.

¹⁾ «Явление Гиббса», ср. [2], гл. IX; см. также {1}, стр. 105.

10. Общий характер сглаживания σ -множителями

Применение σ -множителей к классическим коэффициентам Фурье представляет собой несколько более простой процесс, чем составление средних арифметических частных сумм; он является и менее искажающим с точки зрения точности. Действительность этого процесса сглаживания эквивалентна действительности метода средних арифметических. Можно показать, что специфические свойства ядра Фейера (2.14) сохраняются также и для ядра, которое соответствует σ -процессу. Умножение классических коэффициентов Фурье на множители σ_k оказывает, таким образом, то же влияние на сходимость ряда, что и образование частных сумм; сходимость получается во всех точках, где можно ожидать определенного предела.

В противоположность суммированию с помощью частных сумм, метод σ -множителей является не просто определенным техническим приемом суммирования бесконечного ряда; он имеет самостоятельное значение по отношению к заданной функции $f(x)$. К результату, который дает замена D -процесса $\mathcal{D}_m$ -процессом § 6, можно подойти с совершенно другой точки зрения. Вместо изменения процесса дифференцирования мы теперь изменим рассматриваемую функцию, над которой производится операция, но оставим саму операцию дифференцирования неизменной. Мы можем положить

$$\mathcal{D}_m f(x) = D\bar{f}(x), \quad (4-10.1)$$

где

$$\bar{f}(x) = \frac{m}{2\pi} \int_{-\frac{\pi}{m}}^{+\frac{\pi}{m}} f(x+t) dt. \quad (4-10.2)$$

Это означает, что функция $f(x)$ заменяется новой функцией $\bar{f}(x)$, которая сглаживает первоначальную функцию; значение $\bar{f}(x)$ в каждой точке x равно арифметическому среднему значений $f(x)$ в окрестности $\pm \frac{\pi}{m}$ этой точки (рис. 26). Вместо того чтобы говорить, что

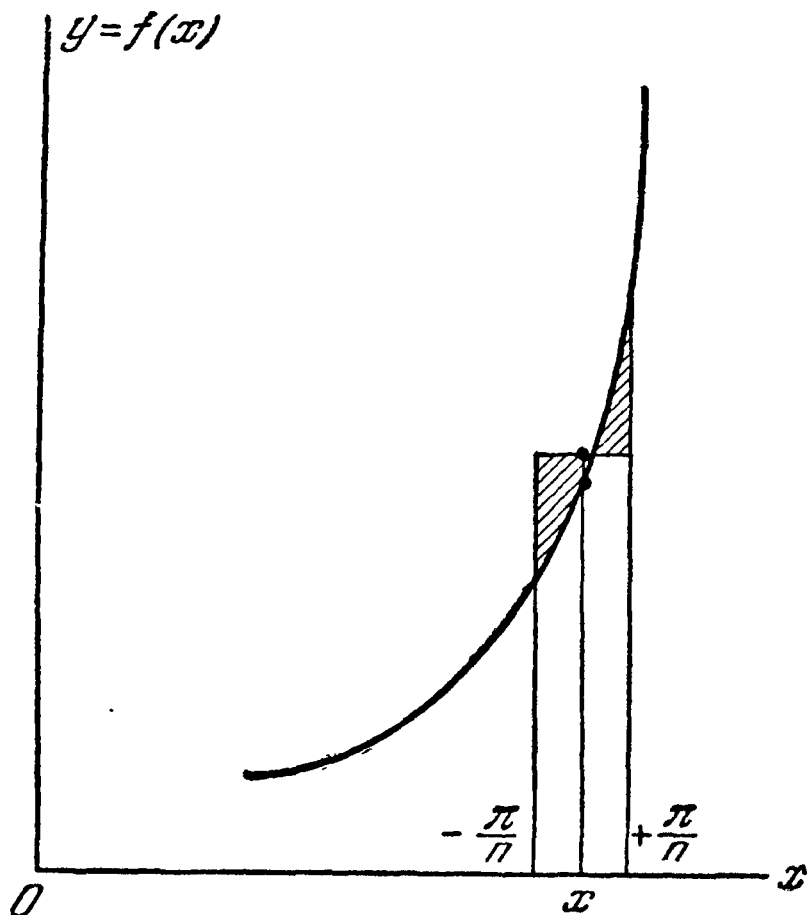


Рис. 26.

коэффициенты усеченного ряда Фурье умножаются на σ_k , мы можем сказать, что операции производятся с усеченным рядом Фурье для измененной функции $\bar{f}(x)$, находящейся в определенном соотношении с заданной функцией $f(x)$, а именно представляющей ее локальную среднюю арифметическую величину. Если m достаточно велико, то локальное усреднение охватывает лишь весьма малую область вокруг точки x , и поэтому искажающее влияние этой операции пренебрежимо мало, за исключением окрестности особой точки, где сглаживание первоначальной неровности функции особенно нежелательно.

Мы видим, таким образом, что за счет сравнительно небольшого искажения функции сходимость ряда Фурье может быть значительно улучшена во всех случаях, когда сходимость первоначального ряда неудовлетворительна. Допущенное при этом искажение можно выразить в регулярной точке функции через вторую производную. Из разложения в ряд Тейлора имеем

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \dots \quad (4-10.3)$$

Рассматривая h как переменную и интегрируя по промежутку $\left(-\frac{\pi}{m}, +\frac{\pi}{m}\right)$, получаем

$$\bar{f}(x) = f(x) + \frac{\pi^2}{6m^2}f''(x). \quad (4-10.4)$$

11. Метод тригонометрической интерполяции

Гармоническому анализу функций часто препятствует трудность вычисления интегралов (2.2). Даже для функций относительно простой структуры могут встретиться при действительном вычислении интегралов (2.2) большие затруднения; интеграл от произведения заданной функции на тригонометрическую функцию типа $\sin nx$ и $\cos nx$ редко выражается через элементарные функции. Поэтому встает вопрос о том, какие меры следует предпринять для практического вычисления коэффициентов Фурье a_k и b_k .

Аналогичные затруднения возникают также в работе с эмпирическими функциями. Значения таких функций заданы лишь для конечной совокупности точек, а именно для точек наблюдения. На практике в большинстве случаев эти точки x_α расположены на равных расстояниях друг от друга. То же самое обстоятельство имеет место для функций, заданных таблично, так как и в этом случае мы имеем в своем распоряжении только значения, вычисленные для равноотстоящих значений независимой переменной x .

Правда, промежуточные точки мы часто можем получать интерполяцией. Это в особенности верно для табулированных функций, которые могут быть вычислены с точностью, достаточной для того, чтобы можно было эффективно интерполировать во всякой точке

заданного интервала. Однако менее благоприятное положение возникает для эмпирических функций, так как здесь часто неизвестен математический закон строения функций; кроме того, «помехи», налагающиеся на истинный ход функции, делают необходимым предварительное сглаживание данных, что не может быть проведено с достаточной степенью надежности. В этих условиях мы избавимся от многих трудностей, если удастся совсем исключить задачу интерполирования, оперируя непосредственно с имеющимся дискретным множеством данных. К счастью, гармонический анализ хорошо приспособлен к случаю равноотстоящих данных. Вычислительный процесс для получения гармонических компонент неизвестной функции, заданной в равноотстоящих точках, прост и непосредствен, и в то же время хорошо сходится. Польза основного оружия прикладного анализа — ряда Фурье — особенно велика потому, что и в случае, когда задано дискретное множество равноотстоящих данных, он оказывается применимым не менее чем в случае непрерывного множества данных.

Общая задача разложения по ортогональным функциям в том случае, когда имеются только дискретные данные, может быть сформулирована следующим образом. Пусть функция $y=f(x)$ задается в дискретных точках $x=x_\alpha$, и значения функции в этих точках обозначены через

$$y_\alpha = f(x_\alpha). \quad (4-11.1)$$

Допустим, что мы аппроксимируем функцию $y=f(x)$ линейной комбинацией данных функций $\varphi_k(x)$:

$$\bar{y} = \sum_{k=1}^m c_k \varphi_k(x). \quad (4-11.2)$$

Примем, что число заданных функций m , вообще говоря, меньше, чем число имеющихся данных n , но не исключим также и предельный случай $m=n$. Мы будем считать, что функции $\varphi_k(x)$ линейно независимы.

Вообще говоря, значения функции $\bar{y}$ в заданных точках x_α не будут совпадать с заданными значениями функции y_α . Определим квадрат средней погрешности нашего приближения, принимая, что он равен сумме квадратов отклонений в данных точках $x=x_\alpha$:

$$\eta^2 = \sum_{\alpha=1}^n (y_\alpha - \bar{y}_\alpha)^2 = \sum_{\alpha=1}^n [y_\alpha - \sum_{k=1}^m c_k \varphi_k(x_\alpha)]^2. \quad (4-11.3)$$

Найдем «наилучшее» приближение, подобрав коэффициенты c_k так, чтобы η^2 было наименьшим. Для этого необходимо, чтобы частные производные от η^2 по c_k равнялись нулю:

$$\sum_{\alpha=1}^n [y_\alpha - \sum_{k=1}^m c_k \varphi_k(x_\alpha)] \varphi_i(x_\alpha) = 0. \quad (4-11.4)$$

Таким образом, мы пришли к системе линейных уравнений для определения c_k . Эта система значительно упрощается, если заданные функции $\varphi_k(x)$ удовлетворяют следующим условиям ортогональности:

$$\sum_{\alpha=1}^n \varphi_i(x_\alpha) \varphi_k(x_\alpha) = 0 \quad (i \neq k). \quad (4-11.5)$$

Эти условия суть прежние условия ортогональности (3.6), но теперь они сформулированы не для процесса интегрирования, а для процесса суммирования. Геометрический термин «ортогональность» связан здесь со следующим понятием аналитической геометрии.

Представим себе некоторое воображаемое евклидово пространство n измерений; значения $\varphi(x_1), \varphi(x_2), \dots, \varphi(x_n)$ произвольной функции $y = \varphi(x)$ рассматриваем как прямоугольные координаты некоторого вектора. Таким образом, функция $\varphi(x)$, взятая только в заданных точках, изображается как некоторый вектор n -мерного пространства. Разлагаемая функция $y = f(x)$ представляет один такой вектор; m функций $\varphi_1(x), \varphi_2(x), \dots, \varphi_m(x)$ представляют m других векторов, которые в силу своей линейной независимости определяют некоторое линейное подпространство m измерений. Приближение (2) можно понимать как произвольный вектор этого линейного подпространства. Метод приведения к минимуму средней ошибки η , т. е. приведения к минимуму суммы (3), геометрически означает приведение к минимуму расстояния от конца данного вектора y до подпространства, построенного на базисных векторах $\varphi_1, \varphi_2, \dots, \varphi_n$. Это означает, что мы *проектируем* y на пространство векторов φ_i *). (См. также V, 16.)

Но процесс проектирования очень упрощается, если базисные векторы φ_i перпендикулярны друг другу, т. е. если «скалярное произведение» любых двух векторов φ_i и φ_k равно нулю:

$$\varphi_i \varphi_k = \sum_{\alpha=1}^n \varphi_i(x_\alpha) \varphi_k(x_\alpha) = 0 \quad (i \neq k). \quad (4-11.6)$$

Мы снова вернулись к нашим предыдущим условиям ортогональности (5), но теперь они интерпретированы в геометрической форме. При выполнении условий ортогональности уравнения (4) упрощаются и принимают вид

$$c_i \sum_{\alpha=1}^n \varphi_i^2(x_\alpha) = \sum_{\alpha=1}^n y_\alpha \varphi_i(x_\alpha). \quad (4-11.7)$$

Таким образом, явное решение поставленной задачи наименьших

*) То есть что вектор $\tilde{y}$, определяемый формулой (2), будет проекцией вектора y на пространство векторов $\varphi_1, \dots, \varphi_n$, если коэффициенты c_k сделать равными решениям системы уравнений (4).

квадратов получает следующий вид:

$$c_i = \frac{y\varphi_i}{\varphi_i\varphi_i} = \frac{\sum_{\alpha=1}^n y_\alpha \varphi_i(x_\alpha)}{N_i}, \quad (4-11.8)$$

где введены «нормы» данных векторов φ_i :

$$N_i = \sum_{\alpha=1}^n \varphi_i^2(x_\alpha). \quad (4-11.9)$$

Если норму каждой функции φ_i , т. е. квадрат длины каждого базисного вектора φ_i нормировать к 1, то мы опять можем назвать нашу систему функций φ_i «ортонормальной». Мы получаем при этом ортогональную систему базисных векторов единичной длины, которая очень удобна для аналитических операций.

Если исходить из такой ортонормальной системы, то прежнее равенство (3.8) принимает следующий вид:

$$\sum_{\alpha=1}^n \bar{y}_\alpha^2 = c_1^2 + c_2^2 + \dots + c_m^2. \quad (4-11.10)$$

Кроме того, сумма квадратов отклонений, которая определяется равенством (11.3), теперь будет равна

$$\sum_{\alpha=1}^n (y_\alpha - \bar{y}_\alpha)^2 = \sum_{\alpha=1}^n y_\alpha^2 - (c_1^2 + c_2^2 + \dots + c_m^2). \quad (4-11.11)$$

Для ненормированных, но ортогональных систем последние два равенства примут следующий вид:

$$\sum_{\alpha=1}^n \bar{y}_\alpha^2 = N_1 c_1^2 + N_2 c_2^2 + \dots + N_m c_m^2, \quad (4-11.12)$$

$$\sum_{\alpha=1}^n (y_\alpha - \bar{y}_\alpha)^2 = \sum_{\alpha=1}^n y_\alpha^2 - (N_1 c_1^2 + \dots + N_m c_m^2). \quad (4-11.13)$$

В предельном случае $m = n$ приближенная функция $\bar{y}$ в заданных точках принимает в точности указанные значения нашей функции y . В этом случае m -мерное подпространство полностью совпадает с n -мерным пространством; оба вектора y и $\bar{y}$ совпадают, и мы имеем

$$y^2 = N_1 c_1^2 + N_2 c_2^2 + \dots + N_n c_n^2. \quad (4-11.14)$$

Если заданные функции $\varphi_k(x)$ имеют комплексные значения, то предыдущие операции нужно несколько видоизменить в том смысле, что φ^2 надо заменить произведением $\varphi\varphi^*$ (см. § 4). В соответствии с

этим в последних трех равенствах c_i^2 надо заменить на $c_i c_i^*$. Условие ортогональности (6) теперь принимает вид $\varphi_i \varphi_k^* = 0$, а формулы (8) для коэффициентов разложения заменяются формулами

$$c_i = \frac{\sum_{\alpha=1}^n y_{\alpha} \varphi_i^*(x_{\alpha})}{N_i}, \quad (4-11.15)$$

$$N_i = \sum_{\alpha=1}^n \varphi_i(x_{\alpha}) \varphi_i^*(x_{\alpha}). \quad (4-11.16)$$

Рассмотрим теперь следующее основное тригонометрическое тождество:

$$\frac{1}{2} e^{-ni\theta} + e^{-(n-1)i\theta} + \dots + e^{(n-1)i\theta} + \frac{1}{2} e^{ni\theta} = \sin n\theta \operatorname{ctg} \frac{\theta}{2} \quad (4-11.17)$$

и примем, что наши функции $\varphi_k(x)$ заданы формулами

$$\varphi_k(x) = e^{ikx} \quad (4-11.18)$$

и что положение заданных точек x выбрано так:

$$x_{\alpha} = \alpha \frac{\pi}{n} \quad [\alpha = -n, -(n-1), \dots, (n-1), n]. \quad (4-11.19)$$

Область изменений переменной x , таким образом, нормируется до промежутка $(-\pi, \pi)$ и делится на $2n$ равных интервалов. Следовательно, число заданных точек равно $2n+1$. В рассматриваемом случае

$$\varphi_j(x_{\alpha}) \varphi_k^*(x_{\alpha}) = e^{i(j-k) \alpha \frac{\pi}{n}}. \quad (4-11.20)$$

Если теперь мы просуммируем эти произведения по α , приняв при этом, что два крайних члена $\alpha = \pm n$ берутся с весом $\frac{1}{2}$, то мы получим левую часть тождества (17), в которой

$$\theta = (j-k) \frac{\pi}{n}. \quad (4-11.21)$$

Так как $j-k$ есть целое число, $\neq 0$, то правая часть формулы (17) становится равной нулю, и мы получаем

$$\varphi_j \varphi_k^* = \sum'_{\alpha=-n}^{+n} \varphi_j(x_{\alpha}) \varphi_k^*(x_{\alpha}) = 0. \quad (4-11.22)$$

Штрих при Σ' означает, что два крайних члена суммы взяты с весом, равным половине.

Нормы функций $\varphi_k(x)$ теперь также просто вычисляются. Полагая $j=k$, получим

$$\varphi_k \varphi_k^* = \sum_{\alpha=-n}^{+n}{}' \varphi_k(x_\alpha) \varphi_k^*(x_\alpha) = 2n. \quad (4-11.23)$$

В заключение мы можем сказать, что решение задачи аппроксимирования по имеющимся данным методом наименьших квадратов с помощью разложения в ряд

$$\bar{y} = \sum_{k=-m}^m c_k e^{ikx} \quad (m \leq n) \quad (4-11.24)$$

представляется в явном виде формулой

$$c_k = \frac{1}{2n} \sum_{\alpha=-n}^n{}' y_\alpha e^{-ikx_\alpha}. \quad (4-11.25)$$

Это же разложение может быть дано более удобно в вещественной форме, если разделить члены, содержащие синус, от членов, содержащих косинусы. Каждая функция может быть записана в виде суммы четной и нечетной функций

$$f(x) = \frac{1}{2}[f(x) + f(-x)] + \frac{1}{2}[f(x) - f(-x)]. \quad (4-11.26)$$

Поэтому достаточно провести рассмотрение для функции $g(x)$ со свойством симметрии

$$g(-x) = g(x) \quad (4-11.27)$$

и функции $h(x)$ со свойством

$$h(-x) = -h(x). \quad (4-11.28)$$

В первом случае $c_{-k} = c_k$, тогда как во втором случае $c_{-k} = -c_k$. Поэтому мы в первом случае получаем

$$g(x) = \frac{1}{2} a_0 + a_1 \cos x + \dots + a_m \cos mx \quad (4-11.29)$$

(последний член берется с весом $\frac{1}{2}$, если $m=n$), где

$$a_k = \frac{2}{n} \sum_{\alpha=0}^n{}' g_\alpha \cos k\alpha \frac{\pi}{n}. \quad (4-11.30)$$

Во втором случае

$$h(x) = b_1 \sin x + \dots + b_m \sin mx, \quad (4-11.31)$$

где

$$b_k = \frac{2}{n} \sum_{\alpha=1}^{n-1} h_\alpha \sin k\alpha \frac{\pi}{n}. \quad (4-11.32)$$

В предельном случае $m=n$ мы располагаем достаточным числом функций, чтобы *точно* представить имеющиеся данные. В этом случае

$\tilde{y}(x)$ будет уже не приближением, а интерполирующей функцией для имеющихся данных. Мы получаем аналитическое выражение в форме тригонометрического полинома наименьшей степени, который точно представляет исходные данные и с некоторой точностью дает промежуточные значения функций. Как велика эта точность, — зависит от заданной функции. Большие возможности тригонометрической интерполяции следуют из того факта, что с возрастанием n функция $\tilde{y}$ аппроксимирует $y(x)$ с постоянно уменьшающимися колебаниями. Для каждой функции с ограниченной вариацией функция, полученная тригонометрической интерполяцией, неограниченно стремится к заданной функции $y = f(x)$ в каждой точке данного интервала, когда число данных точек бесконечно возрастает.

Это свойство тригонометрической интерполяции резко отличается от свойств степенной интерполяции для равноотстоящих данных. Хотя мы всегда можем найти полином $2n$ -й степени, который точно представляет $2n + 1$ значения в равноотстоящих точках данного конечного интервала, колебания ошибки между заданными точками не обязательно будут иметь тенденцию к уменьшению амплитуд при увеличении числа n . В окрестности конца интервала колебания ошибок могут возрасти до бесконечности, давая, таким образом, произвольно большую ошибку *всюду*, кроме заданных точек. Рунге показал, что это имеет место, например, для такой простой и регулярной функции, как

$$y = \frac{1}{1 + 25x^2} \quad (4-11.33)$$

в промежутке $[-1, +1]$ ¹⁾.

Тригонометрическое интерполирование полностью свободно от этого специфического недостатка. Колебания ошибки не имеют тенденции возрастать к концам промежутка, а сохраняют один и тот же порядок величины на всем промежутке. Таким образом, тригонометрическая форма интерполяции как с аналитической, так и с практической точки зрения значительно превосходит полиномиальное интерполирование по заданным значениям, соответствующим равноотстоящим точкам.

12. Интерполяция синусами

Пусть $f(x)$ — нечетная функция. Формула (11.31) используется тогда при $m = n - 1$:

$$h(x) = b_1 \sin x + b_2 \sin 2x + \dots + b_{n-1} \sin (n-1)x, \quad (4-12.1)$$

где

$$b_k = \frac{2}{n} \sum_{\alpha=1}^{n-1} y_{\alpha} \sin k\alpha \frac{\pi}{n}. \quad (4-12.2)$$

¹⁾ Ср. разбор явления Рунге в V, 15.

Член $b_n \sin nx$ нецелесообразно добавлять, так как функция $\sin nx$ равна нулю во всех заданных точках, а это оставляет b_n неопределенным. Действительное число заданных точек фактически не больше, чем $n - 1$, ибо $h(0) = 0$ в силу нечетного характера функции, а значение $h(\pi)$ не может быть выражено формулой (1) в силу того, что все функции $\sin kx$ при $x = \pi$ равны нулю. Ряд (11.31) очень медленно сходится, если $h(x)$ не удовлетворяет краевому условию

$$h(\pi) = 0. \quad (4-12.3)$$

Поэтому применять разложение заданной функции в ряд синусов целесообразно лишь тогда, когда функция равна нулю в точках $x = 0$ и $x = \pi$. Если эти условия не выполняются, то мы можем из заданной функции $f(x)$ вычесть линейный двучлен, положив

$$h(x) = f(x) - (\alpha + \beta x), \quad (4-12.4)$$

где постоянные α и β определяются из условий $h(0) = h(\pi) = 0$, которые дают

$$\left. \begin{aligned} \alpha &= f(0), \\ \beta &= \frac{f(\pi) - f(0)}{\pi}; \end{aligned} \right\} \quad (4-12.5)$$

тогда ряд синусов для $h(x)$ будет удовлетворительно сходиться.

Если все же необходимо найти разложение по синусам для первоначальной функции $f(x)$, то целесообразно действовать указанным выше образом, а затем прибавить известное из теории разложение функции $\alpha + \beta x$ в ряд по синусам.

Вычисление коэффициентов b_k ряда синусов можно проводить следующим образом. Строим матрицу с элементами

$$b_{\alpha\beta} = \sin \alpha\beta \left(\frac{\pi}{n} \right), \quad (4-12.6)$$

выписываем заданные значения y_α в строку

$$(0), y_1, y_2, \dots, y_{n-1}, (0)$$

и умножаем эту строку последовательно на строки матрицы $b_{\alpha\beta}$. Полученные таким образом коэффициенты затем умножаем на постоянную $\frac{2}{n}$; это дает последовательные амплитуды

$$b_1, b_2, \dots, b_{n-1}$$

в разложении по синусам (1).

Построение матрицы (6) упрощается, если сначала сделать ее эскиз в *кодированной форме*. Используем условные числа 1, 2, ..., ..., $n - 1$ и начинаем с «ведущей линии»

$$0, 1, 2, \dots, n - 1, -0, -1, -2, \dots, n - 1, 0. \quad (4-12.7)$$

Эти элементы (числа) мы представляем себе выписанными по окружности так, что последний нуль совпадает с первым. Получаем, таким образом, полный цикл, не имеющий ни начала, ни конца.

Теперь мы выбираем каждый *первый* элемент, затем каждый *второй*, каждый *третий*,... элемент ведущей линии и записываем эти выбранные элементы в последовательные строки матрицы, опуская элемент 0 в начале и в конце строки. Каждая строка содержит $n - 1$ элементов. Например, в случае $n = 6$ получается следующее построение:

ведущая линия:

$$0, 1, 2, 3, 4, 5, -0, -1, -2, -3, -4, -5, 0,$$

кодированная матрица:

$$B_6 = \begin{matrix} & \begin{matrix} y_1 & y_2 & y_3 & y_4 & y_5 \end{matrix} \\ \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 4 & -0 & -2 & -4 \\ 3 & -0 & -3 & 0 & 3 \\ 4 & -2 & 0 & 4 & -2 \\ 5 & -4 & 3 & -2 & 1 \end{pmatrix} \end{matrix}. \quad (4-12.8)$$

Значение этих условных чисел следующее. Мы заменяем условное число k числом $\sin \frac{k\pi}{n}$. Следовательно, в нашем примере ($n = 6$) условные числа должны быть заменены следующими настоящими числами:

$$\begin{aligned} 0 &\rightarrow 0, \\ 1 &\rightarrow \sin 30^\circ = 0,5, \\ 2 &\rightarrow \sin 60^\circ = 0,86603, \\ 3 &\rightarrow \sin 90^\circ = 1, \\ 4 &\rightarrow \sin 120^\circ = 0,86603, \\ 5 &\rightarrow \sin 150^\circ = 0,5. \end{aligned}$$

Матрица-«множитель» (6) теперь построена. Если строку заданных значений $y_1, \dots, y_5$ умножить почленно на последовательные строки матрицы B , то получим 5 чисел $b'_1, \dots, b'_5$; их следует теперь помножить на $\frac{2}{n} = \frac{1}{3}$, и это даст окончательно коэффициенты Фурье

$$b_i = \frac{2}{n} b'_i. \quad (4-12.9)$$

13. Интерполяция косинусами

Допустим теперь, что $f(x)$ есть четная функция. Тогда придется действовать при помощи формулы (11.29) со следующим небольшим изменением: разложение в ряд продолжается до $m = n$, но последний член получает множитель $\frac{1}{2}$:

$$g(x) = \frac{1}{2} a_0 + a_1 \cos x + \dots + a_{n-1} \cos (n-1)x + \frac{1}{2} a_n \cos nx, \quad (4-13.1)$$

где

$$a_k = \frac{2}{n} \sum_{\alpha=0}^n y_{\alpha} \cos k\alpha \frac{\pi}{n}. \quad (4-13.2)$$

Матрица-«множитель» теперь состоит из элементов

$$a_{\alpha\beta} = \cos \alpha\beta \frac{\pi}{n}. \quad (4-13.3)$$

Условная (эскизная) матрица тождественна указанной раньше матрице (12.8) с той только разницей, что теперь столбцы 0 и n нельзя опускать. Нулевой столбец, который соответствует ординате $y_0 = g(0)$, содержит сплошь нули, тогда как столбец n , соответствующий ординате $y_n = g(\pi)$, состоит из элементов 0, —0, 0, —0, ... Кроме того, матрица содержит еще нулевую строку, состоящую сплошь из нулей, которая соответствует коэффициенту $\frac{1}{2} a_0$, отсутствующему в случае разложения по синусам; аналогично имеется теперь n -я строка, которая содержит элементы 0, —0, 0, —0, ...; она соответствует последнему коэффициенту $\frac{1}{2} a_n$. Таким образом, полная матрица множителей A (для случая $n=6$)

$$A_6 = \begin{pmatrix} \frac{1}{2} y_0 & y_1 & y_2 & y_3 & y_4 & y_5 & \frac{1}{2} y_6 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 2 & 3 & 4 & 5 & -0 \\ 0 & 2 & 4 & -0 & -2 & -4 & 0 \\ 0 & 3 & -0 & -3 & 0 & 3 & -0 \\ 0 & 4 & -2 & 0 & 4 & -2 & 0 \\ 0 & 5 & -4 & 3 & -2 & 1 & -0 \\ 0 & -0 & 0 & -0 & 0 & -0 & 0 \end{pmatrix}.$$

Кодовые («условные») числа имеют теперь следующее значение: кодовое число k должно быть заменено числом $\cos \frac{k\pi}{n}$. Таким образом,

$$\begin{aligned} 0 &\rightarrow \cos 0^\circ = 1, \\ 1 &\rightarrow \cos 30^\circ = 0,86603, \\ 2 &\rightarrow \cos 60^\circ = 0,5, \\ 3 &\rightarrow \cos 90^\circ = 0, \\ 4 &\rightarrow \cos 120^\circ = -0,5, \\ 5 &\rightarrow \cos 150^\circ = -0,86603. \end{aligned}$$

После получения коэффициентов a'_k мы умножаем их на $\frac{2}{n}$ для того, чтобы найти коэффициенты a_k :

$$a_k = \frac{2}{n} a'_k.$$

В нашем примере ($n=6$) все коэффициенты a'_k делятся на 3.

Множитель веса $\frac{1}{2}$. Разложение по косинусам имеет особенность, не встречавшуюся нам при разложении по синусам — это множители веса $\frac{1}{2}$ на обоих концах. Надо помнить, что обе крайние ординаты y_0 и y_n входят во все подсчеты с весом $\frac{1}{2}$. Кроме того, оба крайних члена разложения по косинусам также умножаются на $\frac{1}{2}$. Эти особенности не встречаются в разложении по синусам, так как здесь эти два члена равны нулю.

Симметрические свойства матричных множителей A и B . Обе матрицы A и B , определенные формулами (12.6) и (13.3), очевидно, симметричны. Однако они обладают еще некоторыми дополнительными ценными свойствами симметрии, вытекающими из симметрических свойств функций синус и косинус. Мы можем воспользоваться этими свойствами симметрии для уменьшения числа производимых умножений вдвое и даже вчетверо. В связи с этим заметим, что из интерпретации кодовых чисел следует, что элементы k и $n-k$ взаимосвязаны. Число независимых элементов будет не n , а только $\frac{1}{2}n$. Кроме того, столбцы k и $n-k$, а также строки k и $n-k$ взаимосвязаны, отличаясь друг от друга лишь знаком. Если рассматривать отдельно все четные столбцы и все нечетные столбцы и то же самое сделать со строками, то вся матрица уменьшится до четверти своего первоначального размера. Мы разбиваем матрицу на четыре меньшие матрицы, имеющие $\frac{n}{2}$ строк и $\frac{n}{2}$ столбцов, и производим действия только над суммами и разностями ординат y_k и y_{n-k} .

Этим приемом мы уменьшаем вчетверо число умножений, но вынуждены выписывать большее число частных результатов. Мы отделяем коэффициенты четного порядка от коэффициентов нечетного порядка, получая каждый коэффициент как сумму или разность двух частных результатов¹⁾.

Численная проверка. Численная проверка расчетов производится на основании того факта, что суммирование по формулам Фурье с помощью полученных коэффициентов должно дать исходные величины. Ортогональность матриц A и B имеет следствием то, что произведения ординат y_i на матрицы A и B дают соответственно коэффициенты a_k и b_k (если не считать множителя $\frac{2}{n}$), а произведения коэффициентов Фурье a_k на матрицу A или коэффициентов b_k на матрицу B вновь дают первоначальные значения.

14. Гармонический анализ равноотстоящих данных

Рассмотрим периодическую функцию $f(x)$ с периодом $2l$ в интервале $[-l, l]$, значения которой заданы в $2n+1$ равноотстоящих точках

$$x_\alpha = \alpha \frac{l}{h} \quad (\alpha = -n, \dots, +n). \quad (4-14.1)$$

Исследование этой функции может быть выполнено путем образования тригонометрического полинома наименьшего порядка, совпадающего в наблюдаемых точках с заданными значениями функции. В этот полином входят функции

$$\cos k \frac{\pi}{l} x, \quad \sin k \frac{\pi}{l} x. \quad (4-14.2)$$

Они занимают место прежних функций $\cos kx$, $\sin kx$, которые были нами использованы для интервала $[-\pi, +\pi]$. Определение коэффициентов a_k и b_k проводится в точности согласно предыдущему алгоритму и использует ординаты $\frac{1}{2}[f(x_\alpha) + f(-x_\alpha)]$ для разложения по косинусам и ординаты $\frac{1}{2}[f(x_\alpha) - f(-x_\alpha)]$ для разложения по синусам.

Во многих проблемах прикладного анализа исходная функция сама по себе непериодическая, но исследуется в заданном конечном интервале переменной x , скажем, между $x=0$ и $x=l$. Метод тригонометрического разложения может служить средством интерполяции заданных значений функции, так как он дает аналитическое выражение для всей функции $f(x)$, значения которой — табличные или эмпирические — приводятся лишь для дискретного ряда равноотстоящих точек. Такое аналитическое выражение может оказаться важным для

¹⁾ Дальнейшие подробности вычислений и практические примеры см. в [10].

вычисления $f(x)$ в точках, лежащих между заданными точками. Но еще чаще ценность такого аналитического выражения состоит в его *операционных* преимуществах. Многие операции анализа могут быть выполнены сравнительно легко над экспоненциальными функциями, тогда как те же операции над заданными алгебраическими или трансцендентными функциями могут быть только обозначены; действительный численный ответ потребовал бы невыполнимого труда. В таких случаях создаются значительные преимущества, если заданная функция может быть предварительно заменена парэктически *) тригонометрическим разложением, которое хорошо аппроксимирует эту функцию. Именно этой цели часто может служить метод тригонометрической интерполяции.

В таких задачах решающее значение имеет то, что аппроксимирующий ряд должен хорошо сходиться. Для этого необходимо, чтобы функция и первая производная принимали (каждая) одинаковые значения в начале и в конце выбранного периода. Так как это условие обычно не выполняется на обоих концах заданного интервала, то этот интервал не может быть выбран в качестве полного периода гармонического разложения. Если принять интервал $[0, l]$ в качестве полупериода разложения и определить $f(x)$ в отрицательной половине интервала $(-l, l)$ как четную функцию $f(-x) = f(x)$, то мы избегаем разрыва в обоих конечных точках интервала $\pm l$, так как $f(-l) = f(l)$; однако $f'(-l) = -f'(l)$ и, таким образом, мы имеем, вообще говоря, разрыв первой производной. Лучшую сходимость мы получим, если вычтем линейную функцию $\alpha + \beta x$ из заданной функции $f(x)$, оперируя, таким образом, с функцией

$$h(x) = f(x) - (\alpha + \beta x), \quad (4-14.3)$$

которая исчезает в двух концевых точках $x=0$ и $x=l$ [ср. с (12.5)]. Если теперь отразить $h(x)$ как *нечетную* функцию $h(-x) = -h(x)$ и рассматривать $2l$ как полный период гармонического разложения, то будет обеспечена непрерывность функции и ее *первой производной* в двух концевых точках периода, ибо

$$h(-l) = h(l) = 0, \quad h'(-l) = h'(l). \quad (4-14.4)$$

Разложение функции $h(x)$ в ряд по одним синусам

$$h(x) = \sum_{k=1}^{n-1} b_k \sin k \frac{\pi}{l} x, \quad (4-14.5)$$

где b_k получаются согласно § 12, будет теперь удовлетворительно сходиться, так как коэффициенты b_k убывают пропорционально *третьей степени* k .

*) Смысл этого термина разъяснен в предисловии.

Трудно переоценить важность метода тригонометрического интерполирования равноотстоящих данных. Он создает возможность использования мощного орудия гармонического анализа для «трудных функций», для которых коэффициенты Фурье в их первоначальном определении, как определенные интегралы [см. (2.2)], не могут быть вычислены. Метод тригонометрической интерполяции заменяет это интегрирование простым процессом суммирования, производимого над сравнительно небольшим числом равноотстоящих ординат *).

15. Ошибка тригонометрической интерполяции

Не всегда выгодно пользоваться классическими коэффициентами Фурье, как единственной схемой, при разборе характера тригонометрических разложений. Правда, усеченный ряд Фурье (3.1) при всяком значении m является наилучшим приближением в том смысле, что он дает минимальную среднюю квадратичную ошибку. Однако средняя квадратичная ошибка не всегда является лучшим критерием приближения. Кроме того, если коэффициенты этого лучшего приближения могут быть вычислены лишь с большим трудом, то зачастую выгоднее заменить классический ряд другим измененным рядом, который легко вычисляется и ошибка которого тем не менее не существенно хуже, чем первоначальная ошибка. Поэтому полезно сравнить точность тригонометрического ряда, полученного интерполяцией, с точностью усеченного ряда Фурье с тем же числом членов.

Допустим сначала, что заданная функция $f(x)$ не содержит гармонических компонент больше, чем $(n+1)$ членов с косинусами и $(n-1)$ членов с синусами. Тогда заданные $2n+1$ равноотстоящих ординат — их действительное число составляет только $2n$ ввиду граничного условия $f(x_0) = f(x_n)$ — определяют $f(x)$ точно, и ряд, полученный путем тригонометрической интерполяции, совпадает с усеченным рядом Фурье в $2n$ членов, причем оба эти ряда представляют $f(x)$ без всякой ошибки.

Допустим теперь, что заданная функция $f(x)$ содержит дальнейшие гармонические компоненты, и частотный спектр содержит «гармоники» до порядка $2n$ вместо n . Следовательно, $f(x)$ содержит члены, пропорциональные

$$\sin(n+k)x, \quad \cos(n+k)x. \quad (4-15.1)$$

Но коэффициенты Фурье усеченного ряда Фурье не учитывают присутствия этих более высоких гармоник, которые содержат добавки различных частот, давая точный гармонический анализ заданной функции.

*) Из существующих шаблонов, облегчающих вычисления, укажем на [96]; см. также [9a].

Ряд, полученный с помощью тригонометрической интерполяции, ведет себя иначе. Он отзывается на присутствие этих высших гармоник, обращая их в более низкие частоты. Погрешность интерполяционного процесса должна иметь вид

$$\eta_n(x) = (\sin nx) u(x), \quad (4-15.2)$$

так как погрешность равна нулю в заданных точках, т. е. когда $\sin nx$ равно нулю. Но тригонометрические тождества

$$\left. \begin{aligned} \sin(n+k)x + \sin(n-k)x &= 2 \cos kx \sin nx, \\ \cos(n+k)x - \cos(n-k)x &= -2 \sin kx \sin nx \end{aligned} \right\} \quad (4-15.3)$$

показывают, что частота $n+k$ эквивалентна частоте $n-k$ во всех заданных точках. Таким образом, тригонометрическая интерполяция превращает частоту $n+k$ в частоту $n-k$ и учитывает амплитуду этой частоты в полном размере без сдвига фазы в случае ряда по косинусам, в случае же ряда по синусам — амплитуду частоты в полной мере, а фазу со сдвигом в 180° .

Такое же явление «превращения» происходит между частотами $2n$ и $3n$, так как частота $2n+k$ эквивалентна частоте $2n-k = n+(n-k)$ и, следовательно, частоте $n-(n-k) = k$, и т. д.

Это «загрязнение» со стороны более высоких частот показывает, что классические коэффициенты Фурье выгоднее интерполяционных коэффициентов, если нас интересует истинный гармонический спектр заданной функции. Если, однако, нашей целью является просто *аппроксимировать* заданную функцию $f(x)$ гармоническим рядом, то мы не можем заранее сказать, что ошибка усеченного ряда Фурье будет обязательно меньше, чем ошибка интерполяционного ряда. Обе ошибки имеют один и тот же порядок величины, так как функцию погрешности $\eta_n(x)$ интерполяционного ряда можно понимать как спектрально искаженную погрешность усеченного ряда Фурье. Оценка максимальной погрешности одинакова в обоих случаях. *Фактический* максимум погрешности может оказаться более пригодным то для одного, то для другого ряда.

Сравнение ряда Фурье с рядом, полученным тригонометрической интерполяцией, заслуживает более пристального внимания. Указанный выше вывод ряда Фурье, использующий совершенно элементарные средства, как будто указывает на то, что коэффициенты a_k, b_k бесконечного тригонометрического ряда, представляющего $f(x)$, не могут быть ничем иным, как определенными интегралами (2.2). Поэтому господствует представление, что только ряды Фурье дают «точный» результат, тогда как другие ряды, коэффициенты которых получены другим путем, являются лишь «приближенными» по своей природе. Это недоразумение можно отнести за счет того исторического факта, что известный тип предельного перехода получил

в эволюции математики преобладающее значение и поэтому приобрел черты единственности, что ведет к неправильным представлениям.

Когда мы выписываем бесконечный ортонормальный ряд вида

$$f(x) = c_1 \varphi_1(x) + c_2 \varphi_2(x) + \dots \quad (4-15.4)$$

и затем, умножая на $\varphi_k(x)$ и интегрируя почленно, получаем

$$c_k = \int_a^b f(x) \varphi_k(x) dx, \quad (4-15.5)$$

то этот процесс кажется единственным. Важно, однако, иметь в виду *«предпосылки, которые при этом молча подразумеваются»* и которые не были ясны во времена Фурье, но полностью выявились в течение XIX столетия, когда возникло точное понимание понятия предела. Два замечания представляют особый интерес.

Прежде всего надо уяснить себе, что знак равенства в соотношении (4) употреблен не в соответствии с точным его значением и получает здесь новое содержание, которое следует специально указать. Лагранж возражал против утверждения, что неаналитическая функция может быть представлена бесконечной суперпозицией синусов и косинусов, на том основании, что эти функции аналитические, и любое число их в сумме даст также аналитическую функцию. Правильным ответом на это возражение Лангранжа служит то соображение, что знак равенства в бесконечном разложении типа (4) не означает фактического равенства. Он означает лишь, что правая часть по мере прибавления все большего числа членов все ближе подходит к левой части, и что, сложив достаточно большое число членов ряда, мы можем сделать разность между левой и правой частью равенства по абсолютной величине *произвольно малой*, однако *не нулем*. Следовательно, ряды Фурье, подобно всем бесконечным разложениям, дают в этом смысле не точные результаты, но лишь произвольно близкие к ним приближения. Они представляют собой бесконечные *аппроксимирующие процессы*.

Второе замечание касается специального характера бесконечного аппроксимирования, которое достигается с помощью равенства вида (4), т. е. «чисто суммирующего» ряда *). Мы складываем все большее и большее число членов. Это означает, если говорить точно, что следующий процесс переходит от n к $n + 1$. Каждый раз мы добавляем еще одно слагаемое, *ничего не изменяя в предыдущих* слагаемых. Однако это ни в коем случае не является обязательным. Мы могли бы добавлять новое слагаемое к линейной комбинации предыдущих слагаемых, составленной с помощью специально подобранных коэффициентов **). Мы при этом теряли бы в простоте

*) В тексте автора сказано: an equation of «dot-dot-dot type».

**) Более подробно об этом сказано в § 5 гл. VII.

процесса, так как теперь должны были бы вместо одного нового коэффициента использовать $n+1$ новых коэффициентов. Одномерная последовательность коэффициентов заменяется треугольной матрицей коэффициентов. Но эта замена простого сложным может дать простоту в другом отношении. Например, обычно не берущиеся в конечном виде определенные интегралы, необходимые для вычисления коэффициентов Фурье, могут быть заменены легко вычисляемыми суммами, которые требуются при тригонометрическом интерполировании. Кроме того, тригонометрические разложения могут быть распространены на класс функций, который гораздо шире класса абсолютно интегрируемых функций (ср. § 8). Эти ряды отличаются от рядов Фурье характером *аппроксимирования*, но не степенью точности, которая остается «неабсолютной, но произвольно большой» точностью во всех случаях, когда ряды вообще сходятся (ср. также VII, 5).

16. Интерполирование с помощью полиномов Чебышева

Неудобством тригонометрического интерполирования функции $f(x)$, заданной равноотстоящими значениями, является обычно тот факт, что $f(x)$ сама по себе не является периодической функцией x и делается периодической лишь искусственно. Самое большее, на что мы можем практически рассчитывать, — это на непрерывность функции и ее первой производной в начале и конце периода. Прерывность высших производных делает получающийся ряд сравнительно слабо сходящимся. Эту слабую сходимость можно устранить при видоизмененной форме ряда Фурье, которая часто оказывается особенно полезной. Допустим, что интервал определения функции $f(x)$ нормирован до $[-1, +1]$. Пусть $f(x)$ есть аналитическая функция внутри этого интервала и на его концах, но неизвестны какие-либо другие краевые условия.

Перейдем от переменной x к новой переменной θ с помощью преобразования

$$x = \cos \theta. \quad (4-16.1)$$

Заданная функция $f(x)$ переходит теперь в $f(\cos \theta) = \varphi(\theta)$ и таким образом преобразуется в *чисто* периодическую функцию от новой переменной θ .

Если x изменяется от -1 до $+1$, то можно считать, что угол θ изменяется от 0 до π . Так как, однако, замена θ на $-\theta$ оставляет x неизменным, то функция $\varphi(\theta)$ определена и в интервале от $-\pi$ до 0 . Это — четная функция от θ :

$$\varphi(-\theta) = \varphi(\theta). \quad (4-16.2)$$

Кроме того, если $f(x)$ дифференцируема несколько раз по x , то

преобразованная функция $\varphi(\theta)$ дифференцируема столько же раз по θ в каждой точке промежутка $[-\pi, \pi]$, включая границы.

При этих обстоятельствах разложение в ряд Фурье функции $\varphi(\theta)$ будет сходиться гораздо быстрее, чем разложение в ряд Фурье первоначальной функции $f(x)$. Это разложение Фурье имеет лишь члены, содержащие косинусы, так как $\varphi(\theta)$ есть четная функция от θ :

$$\varphi(\theta) = \frac{1}{2} \gamma_0 + \sum_{k=1}^{\infty} \gamma_k \cos k\theta. \quad (4-16.3)$$

Выразим теперь этот ряд через первоначальную переменную x . Выразив функции $\cos k\theta$ через переменную x , мы получим так называемые полиномы Чебышева $T_n(x)$ (ср. V, 20). Они представляют собой попеременно четные и нечетные функции от x и могут быть получены на основе рекуррентной формулы

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) \quad (4-16.4)$$

при начальных условиях $T_0(x) = 1$, $T_1(x) = x$. Полиномы более высоких степеней даны в табл. V.

Разложение (3), выписанное в первоначальной переменной x , приобретает теперь следующий вид:

$$f(x) = \sum_{k=0}^{\infty} a_k T_k(x). \quad (4-16.5)$$

Коэффициенты этого разложения выражаются по обычным формулам — определенными интегралами:

$$a_k = \frac{2}{\pi} \int_0^{\pi} \varphi(\theta) \cos k\theta \, d\theta. \quad (4-16.6)$$

Если перейти к переменной x , то получим

$$a_k = \frac{2}{\pi} \int_{-1}^{+1} f(x) T_k(x) \frac{dx}{\sqrt{1-x^2}}. \quad (4-16.7)$$

Снова можно выдвинуть возражение, что вычисление этих определенных интегралов часто окажется для нас непосильным. Поэтому мы обратимся к методу тригонометрического интерполирования, описанному ранее в § 13. Для этой цели должны быть заданы базисные значения функции y_α в точках

$$\theta_\alpha = \alpha \frac{\pi}{n} \quad (\alpha = 0, 1, \dots, n). \quad (4-16.8)$$

Это означает, что для переменной x заданные точки должны располагаться по закону

$$x_\alpha = \cos \alpha \frac{\pi}{n}. \quad (4-16.9)$$

Это — очень неравномерное распределение точек интерполирования, при котором они накапливаются вблизи двух концевых точек $x = \pm 1$ интервала. Геометрическая интерпретация закона распределения (9) состоит в том, что полуокружность радиуса 1 разделена на n равных частей, и точки деления проектируются на ось x (рис. 27). Равномерное распределение по переменному углу θ_α вызывает очень неравномерное распределение точек-проекций x_α .

Значения функции

$$y_\alpha = f(x_\alpha) = f\left(\cos \alpha \frac{\pi}{n}\right). \quad (4-16.10)$$

должны быть заданы в этих $n+1$ точках, вообще говоря иррациональных. Если не считать этого неудобства, то само интерполирование представляет собой легкий трафаретный процесс, так как коэффициенты вычисляются с помощью матрицы (13.3), приведенной выше в § 13. Мы получаем, таким образом, разложение

$$f(x) = \frac{1}{2} a_0 + a_1 T_1(x) + \dots + \frac{1}{2} a_n T_n(x), \quad (4-16.11)$$

которое теперь можно пересоставить в обыкновенный степенной ряд. Этот ряд сходится гораздо лучше обыкновенного ряда Тейлора. Действительно, в некоторых случаях ряд Тейлора может совсем рас-

ходиться, между тем как сходимость разложения (11) будет обеспечена тем фактом, что всякая непрерывная функция с ограниченной вариацией может быть разложена в равномерно сходящийся ряд Фурье.

Неравномерное распределение заданных точек оказывает в высшей степени благотворное влияние на сходимость полученной интерпо-

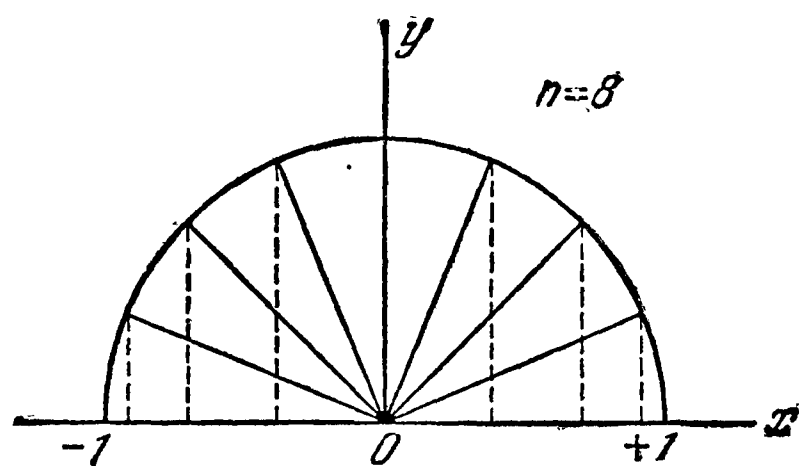


Рис. 27.

ляции. В то время как равноотстоящее интерполирование полиномами дает колебания погрешности, которые сильно возрастают около обоих концов интервала, соответствующее скопление точек задания к двум концевым точкам предупреждает вредное увеличение колебаний*). Погрешность теперь колеблется с амплитудой одного и того же порядка величины во всем интервале. Таким образом, получающийся

*) Подробнее об этом будет сказано в § 11 гл. V.

степенной ряд отличается тем, что он аппроксимирует заданную функцию с меньшей абсолютной максимальной ошибкой, чем аппроксимации, полученные другими полиномами. Кроме того, процесс интерполирования обходится без вычисления определенных интегралов и численно прост. Поэтому рассмотренный здесь метод имеет большое применение на практике. Он переносит ценные аналитические свойства интерполирования тригонометрического типа в область степенных рядов.

17. Интеграл Фурье

В то время как ряды, которые теперь называют по имени Фурье, были уже хорошо известны ко времени Фурье — из основополагающих работ И. Бернулли, Эйлера и Лагранжа, — интеграл Фурье является бесспорно открытием Фурье. Он сделался со временем одним из наиболее мощных средств математического анализа и является особенно важным во всех задачах, касающихся связи между входом и выходом в электрических цепях. Фурье нашел, что разложение произвольных функций на гармонические составляющие остается возможным даже в случае, когда область определения функции $f(x)$ простирается в обе стороны до бесконечности. В этом случае основная частота стремится к нулю и, таким образом, процесс суммирования переходит в интегрирование. Кроме того, пределы интегралов, которые определяют коэффициенты Фурье, из $\pm\pi$ превращаются в $\pm\infty$.

Ряд Фурье есть разложение функции, определенной в конечном интервале, по синусам и косинусам частот, значения которых зависят от заданного интервала. Когда мы распространяем функцию за границы интервала определения, то может быть одно из двух. Либо эта функция математически отображает подлинно периодический процесс, либо она по существу задается лишь в конечном интервале $[-l, +l]$, и мы продолжаем ее периодически для того, чтобы сделать ее представимой рядом Фурье. В последнем случае периодичность функции не вытекает из природы отображаемого процесса, а введена искусственно как математический прием. Если воспользоваться физическим инструментом, вроде волнового анализатора, для определения гармонических составляющих функций, то в первом случае мы фактически получим коэффициенты Фурье как физически измеримые величины. Во втором случае положение совершенно другое. Волновой анализатор не выделяет какой-либо конечный интервал функции $f(x)$. Переменная x теперь представляет время t , анализатор дает информацию о данной функции не только в течение конечного промежутка времени $2l$, а в течение всего времени. Таким образом, встает вопрос, что случится с гармоническим разложением, если основной интервал разложения не ограничен определенной величиной (интервалом определения), а может быть сделан произвольно большим.

Начнем с функции $y=f(x)$, определенной в интервале от $-l$ до $+l$ и удовлетворяющей условиям Дирихле. Этот интервал может

быть преобразован в стандартный интервал $\pm\pi$ путем изменения шкалы: $\xi = \frac{\pi x}{l}$. Выполнив в этом интервале разложение в ряд Фурье, мы можем вернуться к первоначальной переменной x и получить ряд Фурье для произвольного интервала $\pm l$. Воспользуемся математически наиболее удобной комплексной формой записи ряда Фурье [ср. (2.4)]:

$$f(x) = \sum_{k=-\infty}^{+\infty} c_k e^{ik \frac{\pi x}{l}}, \quad (4-17.1)$$

где

$$c_k = \frac{1}{2l} \int_{-l}^{+l} f(x) e^{-ik \frac{\pi x}{l}} dx; \quad (4-17.2)$$

обычная «вещественная» форма записи этого ряда получится, если положить

$$c_k = \frac{1}{2} (a_k - ib_k), \quad c_{-k} = \frac{1}{2} (a_k + ib_k) \quad (4-17.3)$$

и объединить члены с индексами k и $-k$.

Наша функция $f(x)$ была первоначально задана лишь в области $\pm l$; тем не менее мы можем распространить область ее определения до более широкого интервала $\pm L$; вне исходного интервала определим функцию $f(x)$, приписав ей значения, равные нулю (рис. 28).

Новая функция $y = \tilde{f}(x)$, определенная в расширенном интервале, снова может быть разложена в ряд Фурье. Заменяя в формулах (1) и (2) l через L , получим

$$\tilde{f}(x) = \sum_{k=-\infty}^{+\infty} \tilde{c}_k e^{\frac{ik\pi x}{L}}, \quad (4-17.4)$$

$$\tilde{c}_k = \frac{1}{2L} \int_{-L}^{+L} \tilde{f}(x) e^{-ik \frac{\pi x}{L}} dx. \quad (4-17.5)$$

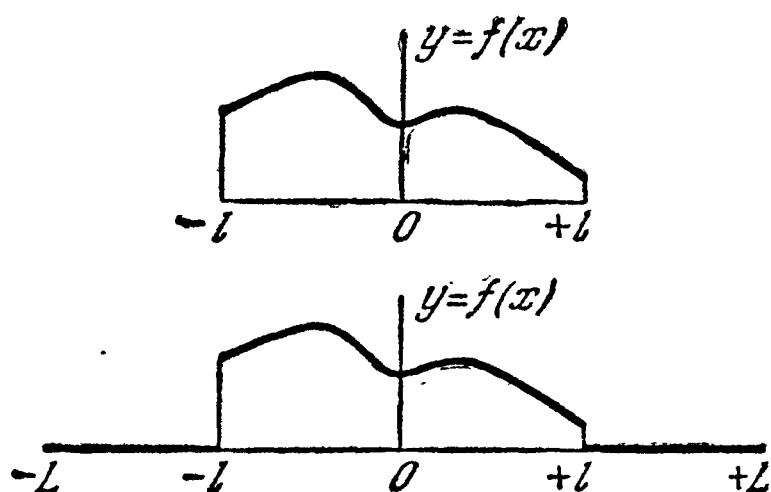


Рис. 28.

Хотя новый ряд (4) есть разложение функции $f(x)$ по совершенно другим частотам с совершенно новыми коэффициентами, он тем не менее

дает приближенные значения прежней функции $f(x)$ для любого значения x в интервале $\pm l$, тогда как вне этого интервала (вплоть до границ $\pm L$) новый ряд дает значения, близкие к нулю. Рассмотрим подробнее гармонические составляющие функции $f(x)$. Если гармоническая функция записана в форме $\sin 2\pi \nu t$ или $\cos 2\pi \nu t$, то мы назовем ν — число колебаний в секунду — «частотой» гармонических колеба-

ний. Гармонические функции, входящие в рассмотренный ранее ряд (1), можно записать в виде

$$\cos k \frac{2\pi x}{2l} + i \sin k \frac{2\pi x}{2l}.$$

Следовательно, в этом первоначальном разложении (1) имеются частоты

$$\nu_1 = \frac{1}{2l}, \quad \nu_2 = \frac{2}{2l}, \quad \dots, \quad \nu_k = \frac{k}{2l}, \quad \dots$$

Соответствующими частотами в нашем новом разложении (4) будут

$$\nu_1 = \frac{1}{2L}, \quad \nu_2 = \frac{2}{2L}, \quad \dots, \quad \nu_k = \frac{k}{2L}, \quad \dots$$

Если L в два раза больше, чем l , то новая основная частота будет равна *половине* прежней основной частоты. Следовательно, гармоники 1, 2, 3, 4... первого разложения становятся теперь гармониками 2, 4, 6, 8... второго разложения, и наше новое разложение в ряд Фурье содержит вдвое больше частот, чем прежнее.

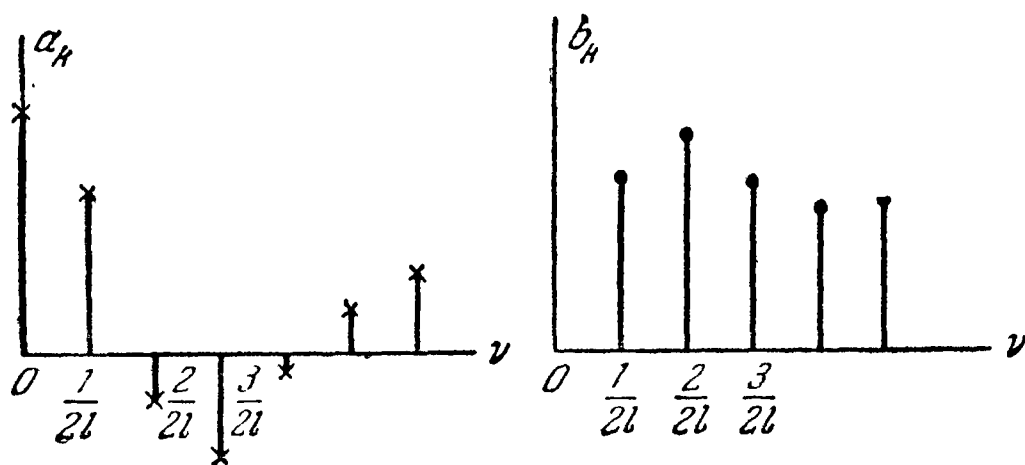


Рис. 29.

Чтобы изучить распределение гармонических составляющих, мы представим их графически. В качестве абсциссы мы выбираем частоту, которой они соответствуют. Следовательно, графическая картина гармонического разложения (1), если представить его в вещественной форме, имеет вид, изображенный на рис. 29. Если такое же графическое представление сделать для разложения (4), то мы получим больше отрезков, так как имеется больше частот. Условимся опускать в формуле (5) постоянный множитель $\frac{1}{2}L$ и изображать на графике только величины

$$\gamma_k = \int_{-l}^{+l} \tilde{f}(x) e^{\frac{-2\pi i k x}{2L}} dx. \quad (4-17.6)$$

Тогда

$$\tilde{c}_k = \frac{\gamma_k}{2L}; \quad (4-17.7)$$

амплитуды γ_k имеют то преимущество, что их не надо изменять при увеличении L ; нам просто надо пополнять чертеж все новыми и новыми ординатами. Таким образом, как нетрудно заметить, наш график просто отбирает ряд определенных ординат некоторой универсальной функции. Введем в рассмотрение следующую функцию непрерывной переменной ν :

$$F(\nu) = \int_{-l}^{+l} \tilde{f}(x) e^{-2\pi i \nu x} dx. \quad (4-17.8)$$

Эта функция содержит все возможные гармонические составляющие функции $\tilde{f}(x)$ независимо от того, насколько малым или большим является L , причем

$$\gamma_k = F\left(\frac{k}{2L}\right). \quad (4-17.9)$$

Наш прежний график теперь заменяется следующим (см. рис. 30). Замечательный факт состоит в том, что какой бы нерегулярной ни была

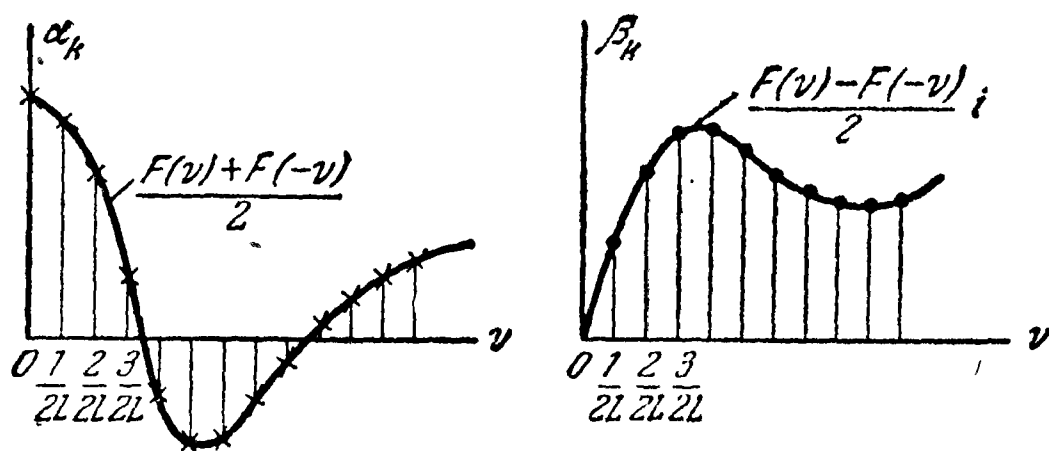


Рис. 30.

функция $f(x)$, новая функция $F(\nu)$ всегда является *непрерывной* и даже *аналитической* (т. е. неограниченно дифференцируемой). Исходная функция $f(x)$ может иметь любое число разрывов или других особенностей, но новая функция $F(\nu)$ будет аналитической для всех (вещественных или комплексных) значений ν . Она называется «преобразованием Фурье» первоначальной функции $f(x)$. Функция $F(\nu)$ совсем не похожа на первоначальную функцию; она просто *соответствует* ей, до некоторой степени аналогично тому, как логарифм числа соответствует самому числу.

Посмотрим теперь, что случится, когда мы будем придавать L все большие и большие значения, постепенно увеличивая период нашего разложения до бесконечности. Тогда ординаты, которые нам надо вычерчивать на графике, располагаются все гуще и гуще и в пределе, когда L возрастает до бесконечности, исчезает выделение каких-либо специфических частот, *все частоты оказываются равноправно представленными на графике*. Прежний линейчатый спектр превратился

в пределе в сплошной спектр, и новая функция $F(\nu)$ в целом даст теперь распределение гармонических амплитуд функции $f(x)$.

Что можно теперь сказать относительно синтеза Фурье, выраженного в форме ряда Фурье? Этот ряд находится в очень тесной связи с получившейся в результате преобразования функцией $F(\nu)$. Формулы (4) и (5), если использовать функцию $F(\nu)$, объединяются в одну формулу

$$f(x) = \frac{1}{2L} \sum_{k=-\infty}^{+\infty} F\left(\frac{k}{2L}\right) e^{\frac{2\pi i k x}{2L}}. \quad (4-17.10)$$

Определим теперь следующую функцию частоты ν :

$$G(\nu) = F(\nu) e^{2\pi i \nu x}; \quad (4-17.11)$$

тогда

$$f(x) = \frac{1}{2L} \sum_{k=-\infty}^{+\infty} G\left(\frac{k}{2L}\right), \quad (4-17.12)$$

или, полагая

$$\frac{1}{2L} = \varepsilon, \quad (4-17.13)$$

получаем

$$f(x) = \varepsilon \sum_{k=-\infty}^{+\infty} G(k\varepsilon). \quad (4-17.14)$$

На рис. 31 изображен график функции $G(\nu)$. Этот рисунок сделан в предположении, что x — заданная константа; то, что значения

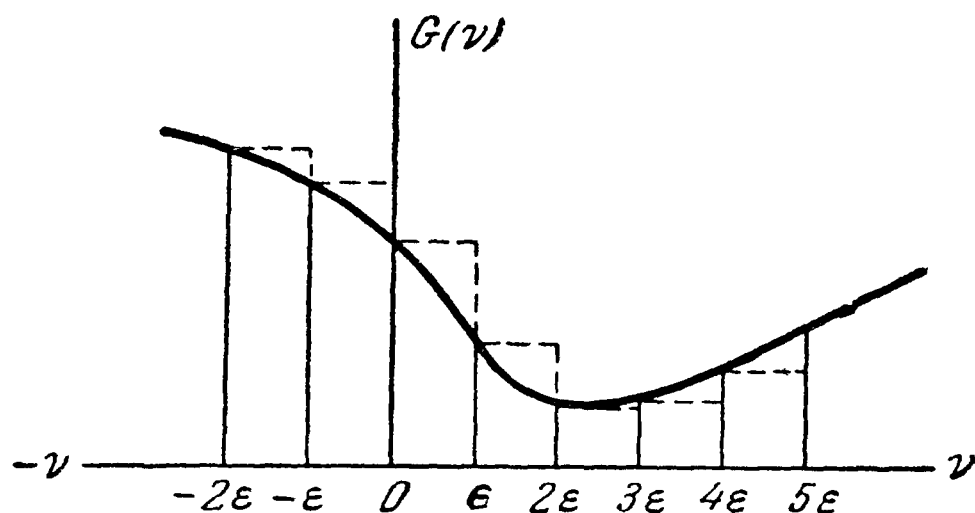


Рис. 31.

функции $G(\nu)$ комплексны, также не принято во внимание в целях наглядности.

Чем больше основной интервал $2L$, тем гуще расположатся ординаты $G(\nu)$, участвующие в сумме Фурье. Допустим теперь, что L возрастает до бесконечности. Это означает, что ε стремится к нулю. По основной теореме интегрального исчисления сумма (14) с прибли-

жением ε к нулю стремится к определенному пределу. Этот предел есть *полная площадь*, ограниченная кривой $y = G(\nu)$:

$$f(x) = \int_{-\infty}^{+\infty} G(\nu) d\nu = \int_{-\infty}^{+\infty} F(\nu) e^{2\pi i \nu x} d\nu. \quad (4-17.15)$$

Сумма Фурье все более и более приближается к *интегралу* Фурье. В то время как преобразование Фурье $F(\nu)$ функции $f(x)$ разлагает заданную функцию $f(x)$ на ее гармонические составляющие, интеграл Фурье *синтезирует* эти гармонические составляющие в первоначальную функцию. Обратим внимание на замечательную двойственность двух равенств:

$$F(\nu) = \int_{-l}^{+l} f(x) e^{-2\pi i \nu x} dx,$$

$$f(x) = \int_{-\infty}^{+\infty} F(\nu) e^{2\pi i \nu x} d\nu.$$

Единственное различие между ними состоит в том, что в первом интеграле пределами интегрирования служат $\pm l$, а во втором $\pm \infty$. Но это происходит лишь потому, что до сих пор мы считали функцию $f(x)$ определенной только между $\pm l$ и равной нулю вне этого интервала. Мы можем расширить область определения функции $f(x)$ до произвольно большого интервала и постепенно приблизиться к бесконечности, не нарушая правильности наших результатов. В пределе мы получаем

$$F(\nu) = \int_{-\infty}^{+\infty} f(x) e^{-2\pi i \nu x} dx, \quad (4-17.16)$$

$$f(x) = \int_{-\infty}^{+\infty} F(\nu) e^{2\pi i \nu x} d\nu. \quad (4-17.17)$$

Взаимность теперь полная. Единственное, что мы теряем при рассмотрении бесконечных пределов интегрирования в первой формуле, это то, что $F(\nu)$ — уже больше не аналитическая и даже не обязательно непрерывная функция от ν . Переменная ν принимает только вещественные значения. Но функция $F(\nu)$ все же вполне определится, если только $f(x)$ абсолютно интегрируема, т. е. если существует интеграл

$$\int_{-\infty}^{+\infty} |f(x)| dx \quad (4-17.18)$$

и $f(x)$ имеет ограниченную вариацию в каждом конечном интервале.

Для того, чтобы отчетливо представить себе характеристические сходство и различие между рядом Фурье и интегралом Фурье, мы еще раз сопоставим основные формулы:

ряд Фурье; дискретный спектр (частоты $\frac{1}{2l}, \frac{2}{2l}, \dots, \frac{k}{2l}$):

$$f(x) = \sum_{k=-\infty}^{+\infty} c_k e^{2\pi i k \left(\frac{x}{2l}\right)}, \quad c_k = \frac{1}{2l} \int_{-l}^{+l} f(x) e^{-2\pi i k \left(\frac{x}{2l}\right)} dx;$$

интеграл Фурье; непрерывный спектр (все частоты между $-\infty$ и $+\infty$):

$$f(x) = \int_{-\infty}^{+\infty} F(\nu) e^{2\pi i \nu x} d\nu, \quad F(\nu) = \int_{-\infty}^{+\infty} f(x) e^{-2\pi i \nu x} dx.$$

Вопрос: Что означает отрицательная частота?

Ответ: Если мы оперируем только с положительными частотами, то каждой частоте ν соответствуют две функции, а именно $\sin 2\pi \nu x$ и $\cos 2\pi \nu x$. Этого удвоения можно избежать, введя положительные и отрицательные частоты и поставив в соответствие каждому значению ν единственную функцию $e^{2\pi i \nu x}$. Тогда произвольное реальное колебание является всегда взаимодействием двух частот $+\nu$ и $-\nu$.

При практическом использовании преобразования Фурье зачастую удобнее ввести «угловую частоту» $\omega = 2\pi \nu$ вместо обыкновенной частоты ν и записывать формулы гармонического анализа и синтеза в следующей форме¹⁾ *):

$$\begin{aligned} F(\omega) &= \int_{-\infty}^{+\infty} f(x) e^{-i\omega x} dx, \\ f(x) &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} F(\omega) e^{i\omega x} d\omega. \end{aligned} \quad (4-17.19)$$

¹⁾ По причинам исторического характера почти во всей математической литературе теория интеграла Фурье рассматривается на основе так называемого «двойного интеграла Фурье», который объединяет гармонический анализ и синтез в одну операцию. Подобный подход настолько затемняет результат, что теория интеграла Фурье, несмотря на его фундаментальную важность в обширных областях физики и техники, часто представляется как одна из наиболее туманных и наименее понятных глав высшего анализа.

*) Строгий вывод формул (19), не основывающийся на теории «двойного интеграла Фурье», а использующий только такие же элементарные средства, на которых базировался вышеизложенный вывод, читатель найдет в статье И. М. Гельфанда, О формуле преобразования Фурье, Математическое просвещение, вып. 5, стр. 155—159 (1960).

18. Соотношение входа и выхода электрических цепей

Интеграл Фурье является одним из наиболее мощных средств прикладного анализа. Он играет фундаментальную роль во всех проблемах электрических цепей, но область применения его гораздо шире, так как условия, характерные для электрической цепи, встречаются также во многих других задачах физики и техники.

Эти общие условия могут быть описаны следующим образом. Имеется измерительный прибор типа гальванометра, который регистрирует некоторую заданную функцию времени t . Мы имеем, таким образом, две функции, а именно «функцию входа», или «сигнал», $f(t)$ и «функцию выхода», или «ответ», $g(t)$. Измерительным прибором может быть телефон или громкоговоритель, или другое средство связи. Это может быть также какой-либо сервомеханизм, отвечающий на заданную «команду». Его роль может быть описана в задаче с краевым условием, в которой правая часть уравнения в частных производных есть вход, а решение уравнения — выход. Во всех этих случаях общая схема действия может быть охарактеризована некоторыми чертами, свойственными целой группе задач.

Имеются две функции $f(t)$ и $g(t)$ — вход и выход. Заданный физический механизм определяет соотношение между этими двумя функциями. Мы можем представить себе $g(t)$ как некоторого рода «отображение» функции $f(t)$ и будем говорить, что это отображение имеет «тип C», если выполняются следующие общие условия.

1. Отображение линейно. Это значит, что если $f(t)$ переходит в $\alpha f(t)$, то $g(t)$ переходит в $\alpha g(t)$. Кроме того, если $f(t)$ есть линейная суперпозиция некоторого числа функций, т. е.

$$f(t) = \alpha_1 f_1(t) + \alpha_2 f_2(t) + \dots + \alpha_n f_n(t), \quad (4-18.1)$$

то $g(t)$ есть такая же суперпозиция соответствующих отображенных функций

$$g(t) = \alpha_1 g_1(t) + \alpha_2 g_2(t) + \dots + \alpha_n g_n(t). \quad (4-18.2)$$

2. Если $f(t)$ есть периодическая функция, записанная в комплексной форме,

$$f(t) = e^{i\omega t}, \quad (4-18.3)$$

то $g(t)$ воспроизводит эту функцию с простым множителем пропорциональности,

$$g(t) = \varphi(\omega) e^{i\omega t}. \quad (4-18.4)$$

Множитель $\varphi(\omega)$, вообще говоря, комплексный и поэтому может быть разбит на вещественную и мнимую части:

$$\varphi(\omega) = A(\omega) + i B(\omega). \quad (4-18.5)$$

Множитель $\varphi(\omega)$ называется «передаточной функцией». Физический смысл этой функции состоит в том, что если *вход* есть строго периодическая функция определенной частоты и постоянной амплитуды, то *выход* есть также периодическая функция той же частоты, но другой амплитуды и другой фазы. Изменение амплитуды и фазы зависит от частоты и характеризуется множителем $\varphi(\omega)$. Реакция $g(t)$ на строго синусоидальную функцию входа называется «установившейся реакцией» или «частотной реакцией».

Интеграл Фурье разлагает произвольную функцию $f(t)$ на ее гармонические составляющие [ср. (17.19)]:

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} F(\omega) e^{i\omega t} d\omega. \quad (4-18.6)$$

В силу принципа суперпозиции прибор связи воспринимает каждую из этих гармонических составляющих и налагает на нее свою собственную передаточную функцию. Затем гармонические составляющие вновь синтезируются. В результате этой операции мы получаем:

$$g(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} F(\omega) \varphi(\omega) e^{i\omega t} d\omega. \quad (4-18.7)$$

Итак, как видно, для всякой заданной $f(t)$ можно найти соответствующую функцию $g(t)$, если задана передаточная функция $\varphi(\omega)$.

Однако соотношение между $f(t)$ и $g(t)$ может быть выражено и в совершенно другой форме, если ввести вторую фундаментальную функцию, с помощью которой может быть охарактеризовано действие цепи. Мы видели [ср. (17.19)], что по теореме взаимности Фурье соотношение (18.6) обратимо:

$$F(\omega) = \int_{-\infty}^{+\infty} f(t) e^{-i\omega t} dt. \quad (4-18.8)$$

Функцию $F(\omega)$ мы можем ввести в равенство (7) и записать результат в следующем виде ¹⁾:

$$g(t) = \int_{-\infty}^{+\infty} f(\tau) K(t - \tau) d\tau, \quad (4-18.9)$$

¹⁾ Это преобразование не очевидно, ибо оно содержит двойной предельный переход. Однако его законность может быть обоснована для всех функций с ограниченной вариацией.

где

$$K(\xi) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \varphi(\omega) e^{i\omega\xi} d\omega. \quad (4-18.10)$$

Но мы знаем из самой природы «реакций», что она не может предшествовать сигналу, а должна *следовать* за ним. Будущее не может влиять на прошлое. Следовательно, $g(t)$ не может зависеть от значений функции $f(t)$, следующих за моментом времени $t = \tau$. По этим соображениям мы должны потребовать для всех «устойчивых» отображений типа C выполнения следующего дополнительного условия:

$$K(t - \tau) = 0 \quad \text{при} \quad \tau > t,$$

которое означает, что

$$K(\xi) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \varphi(\omega) e^{i\omega\xi} d\omega = 0 \quad (\xi < 0). \quad (4-18.11)$$

В этом случае верхний предел интеграла (9) будет t :

$$g(t) = \int_{-\infty}^t f(\tau) K(t - \tau) d\tau. \quad (4-18.12)$$

Физический сигнал $f(t)$ обыкновенно начинается не при $t = -\infty$, а в определенный момент времени, который может быть нормирован так, чтобы $t = 0$ был начальным моментом. Тогда функция выхода $g(t)$ принимает вид

$$g(t) = \int_0^t f(t) K(t - \tau) d\tau. \quad (4-18.13)$$

Основной величиной, характеризующей соотношение вход — выход для данного прибора, является при этом новом представлении функция $K(\xi)$. Она имеет вполне определенный физический смысл. Допустим, что мы используем сигнал $f(t)$, который длится лишь в течение небольшого промежутка времени между $t = 0$ и $t = \varepsilon$. В течение этого промежутка времени функция $f(t)$ должна скачком подняться от 0 до большого постоянного значения $1/\varepsilon$ и затем опять упасть до значения 0. Сделаем теперь величину ε произвольно малой (рис. 32). Такого рода сигнал (сравнимый с ударом молота) называется «единичным импульсом», а соответствующая реакция — «импульсной реакцией».

Так как $f(t)$ теперь рассматривается только в бесконечно малом промежутке времени, в течение которого функция $K(t-\tau)$ практически остается постоянной, то мы можем записать формулу (13) в виде

$$g(t) = K(t) \int_0^\epsilon f(\tau) d\tau. \quad (4-18.14)$$

Это показывает, что функция $K(t)$ имеет значение «импульсной реакции» прибора¹⁾.

Согласно формуле (10) импульсная реакция и реакция установившегося состояния находятся в определенной связи друг с другом

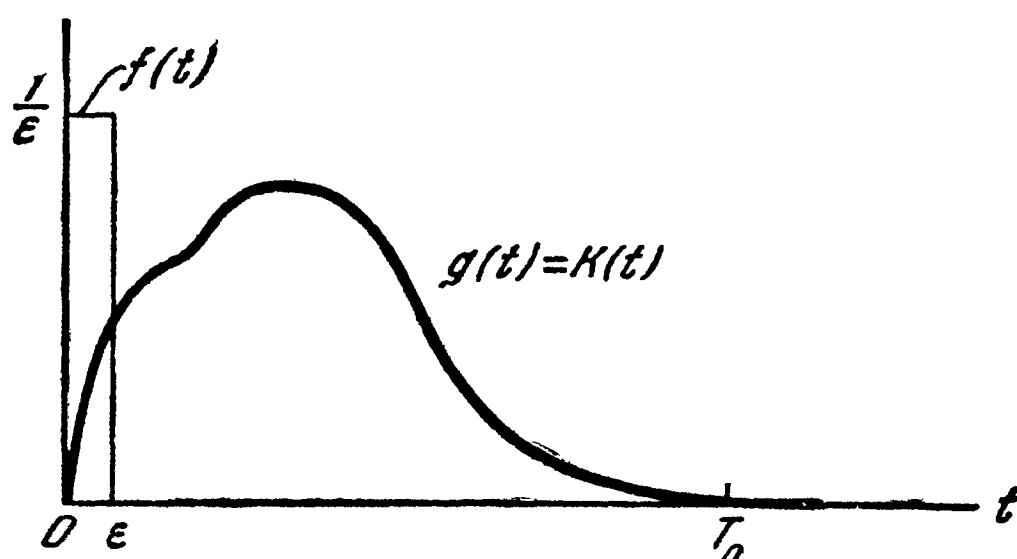


Рис. 32.

одна является преобразованием Фурье другой). Если передаточная функция $\varphi(\omega)$ задана, то мы можем получить $K(t)$ на основе интеграла Фурье (10). Но тогда по теореме взаимности Фурье мы можем также положить

$$\varphi(\omega) = \int_0^\infty K(t) e^{-i\omega t} dt \quad (4-18.15)$$

(нижним пределом служит нуль, так как $K(t)$ равно нулю при отрицательном t). Отделяя вещественную и мнимую части согласно равенству (5), получаем

$$\left. \begin{aligned} A(\omega) &= \int_0^\infty K(t) \cos \omega t dt, \\ B(\omega) &= - \int_0^\infty K(t) \sin \omega t dt. \end{aligned} \right\} \quad (4-18.16)$$

¹⁾ По причинам исторического характера электрики часто предпочитают характеризовать переходный процесс в цепи сети «реакцией на единичный толчок». Два вида реакции находятся между собой в простом отношении, поскольку импульсная реакция есть производная по времени от реакции на единичный толчок.

Это показывает, что $A(\omega)$ — всегда *четная*, а $B(\omega)$ — всегда *нечетная* функция от ω :

$$A(-\omega) = A(\omega); \quad B(-\omega) = -B(\omega). \quad (4-18.17)$$

Возвращаясь к равенству (10) и записав его в вещественной форме, находим

$$K(t) = \frac{1}{\pi} \left[\int_0^{\infty} A(\omega) \cos \omega t d\omega - \int_0^{\infty} B(\omega) \sin \omega t d\omega \right]. \quad (4-18.18)$$

Однако условие устойчивости (11) устанавливает определенную зависимость между $A(\omega)$ и $B(\omega)$, а именно для всех положительных значений t должно иметь место равенство

$$\int_0^{\infty} A(\omega) \cos \omega t d\omega = - \int_0^{\infty} B(\omega) \sin \omega t d\omega. \quad (4-18.19)$$

Следовательно, достаточно знать *либо* вещественную, *либо* мнимую часть передаточной функции:

$$K(t) = \frac{2}{\pi} \int_0^{\infty} A(\omega) \cos \omega t d\omega = - \frac{2}{\pi} \int_0^{\infty} B(\omega) \sin \omega t dt. \quad (4-18.20)$$

19. Эмпирическое определение соотношения входа — выхода

Часто оказывается возможным определить частоту реакции механизма путем прямых наблюдений. Мы возбуждаем в механизме вынужденные колебания, пользуясь синусоидальной входной функцией и выжидая, пока пройдет неустановившаяся реакция. Если затем мы измерим амплитуду выхода и сдвиг фазы между этими двумя функциями, то получим $\varphi(\omega)$ для данной частоты ω . Измерения следует проводить в достаточно широком диапазоне частот.

Во многих случаях более удобно пользоваться импульсом в качестве входа. Наблюдая выход, мы узнаем функцию $K(t)$, а преобразование Фурье этой функции даст $\varphi(\omega)$. Трудность состоит лишь в том, что реализация очень резкого «удара» часто физически невозможна без нанесения серьезных повреждений данному прибору. Приходится идти окольным методом. Воспользуемся некоторой наблюдаемой функцией $f(t)$ как входом, и получим наблюдаемую функцию $g(t)$ как выход. Из этих данных нужно получить путем вычислений импульсную реакцию $K(t)$ или частотную реакцию $\varphi(\omega)$. Хотя мы оставляем характер функции $f(t)$ произвольным, мы все же принимаем, что она имеет одно общее свойство, которым она походит на импульс: $f(t)$ начинается с нуля при $t=0$ и сходит снова на нуль после некоторого конечного промежутка времени t_0 .

Есть еще одна черта, свойственная импульсной реакции, которую мы не рассматривали в наших рассуждениях. Все приборы, которыми пользуются в системах связи, а также сервомеханизмы обладают свойством рассеивать энергию, полученную ими от входного импульса. Поэтому $K(t)$ становится практически равной нулю после некоторого времени T_0 , которое мы назовем «временем запоминания» прибора¹⁾. Так как мы приняли, что функция $f(t)$ исчезает после промежутка времени t_0 , то теперь мы должны дополнительно принять, что выход $g(t)$ практически исчезает после промежутка времени

$$T_1 = t_0 + T_0 \quad (4-19.1)$$

(см. рис. 33). Представим себе теперь, что одна и та же функция $f(t)$ подается на вход снова и снова до и после $t=0$ через интервалы T_1 , т. е. в начальные моменты времени $t=0, \pm T_1, \pm 2T_1, \dots$

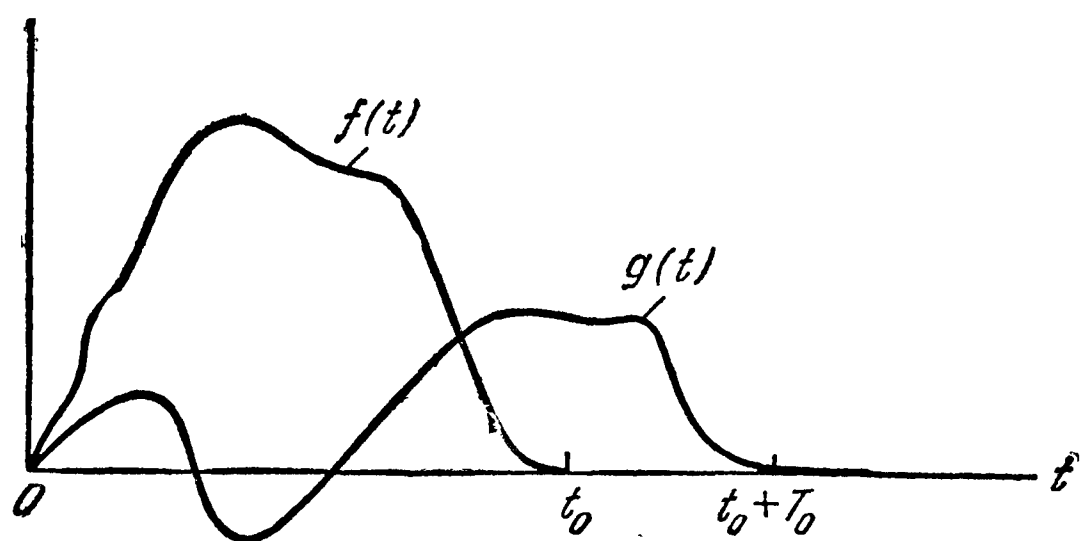


Рис. 33.

Таким образом, $f(t)$ представляет собой в этом случае периодическую функцию с периодом T_1 , и это же имеет место для $g(t)$. Внутри промежутка времени T_1 никакого изменения этих функций не происходит, ибо мы ждали, пока $g(t)$ сделается равной нулю. Теперь мы можем при помощи ряда Фурье разложить и $f(t)$ и $g(t)$ по гармоникам в интервале $(0, T_1)$:

$$f(t) = \sum_{k=-\infty}^{+\infty} c_k e^{\frac{ik2\pi t}{T_1}}, \quad g(t) = \sum_{k=-\infty}^{+\infty} \bar{c}_k e^{\frac{ik2\pi t}{T_1}}. \quad (4-19.2)$$

Практически в обоих разложениях будет лишь конечное число членов, и коэффициенты c_k и $\bar{c}_k$ лучше определять тригонометрической интерполяцией, чем интегрированием (ср. §§ 11—13). Наша входная

¹⁾ Время запоминания не является абсолютно точно определенной величиной, ибо $K(t)$ становится равной нулю лишь асимптотически. Но при заданной точности существует известное T_0 , после которого $K(t)$ становится пренебрежимо малой. Чем больше требуемая точность, тем больше время запоминания.

функция теперь есть суперпозиция строго синусоидальных функций, и это же можно сказать о выходе. По определению передаточной функции $\varphi(\nu)$, мы имеем

$$\varphi\left(\frac{k}{T_1}\right) = \frac{\bar{c}_k}{c_k} \quad (k=0, \pm 1, \pm 2, \dots). \quad (4-19.3)$$

Мы получаем, таким образом, значения $\varphi(\nu)$ лишь для дискретного ряда равноотстоящих точек; для других значений ν можно найти $\varphi(\nu)$ при помощи линейной (или квадратичной) интерполяции.

Все же остается неразрешенным то затруднение, что заданная функция $f(t)$ могла быть слишком гладкой для наших целей. Если $f(t)$ не содержит достаточного количества гармоник обертонов, то определение отношения (3) при k , большем некоторого значения, может оказаться невозможным ввиду малости знаменателя. Тогда придется остановиться на значении $\varphi(\nu)$, которое еще не пренебрежимо мало. Однако для нас достаточно установления *асимптотического закона* для функции $\varphi(\nu)$. Скорость, с которой $\varphi(\nu)$ уменьшается до нуля, по меньшей мере равна ν^{-1} . Если последние несколько наблюденных значений $\varphi(\nu)$ согласуются со схемой

$$\varphi(\nu) = \frac{ia_1}{\nu}, \quad (4-19.4)$$

или, более точно, со схемой

$$\varphi(\nu) = \frac{ia_1}{\nu} + \frac{a_2}{\nu^2}, \quad (4-19.5)$$

то мы можем быть удовлетворены, ибо для всех частот вне интервала наших измерений значение $\varphi(\nu)$ можно получить экстраполяцией.

Если «установившаяся реакция» была получена таким образом, то импульсная реакция также может быть найдена на основе тех же данных. Воспользуемся приемом, при котором импульсный вход повторяется через правильные интервалы T_0 . Ряд Фурье для этого импульса («дельта-функция») есть расходящийся ряд

$$f(t) = \frac{1}{T_0} \sum_{k=-\infty}^{+\infty} e^{\frac{2\pi i k t}{T_0}}. \quad (4-19.6)$$

Этот ряд, хотя и бесполезный сам по себе, становится сходящимся, если применить передаточную функцию к каждой имеющейся частоте

$$K(t) = \frac{1}{T_0} \sum_{k=-\infty}^{+\infty} \varphi\left(\frac{k}{T_0}\right) e^{\frac{2\pi i k t}{T_0}}. \quad (4-19.7)$$

Хотя этот бесконечный ряд и сходится, его сходимость может оказаться слишком слабой для практических целей, особенно если

асимптотическая схема функции $\varphi(\nu)$ принадлежит к типу (4). Мы ускорим сходимость, полагая

$$K(t) = A_1 e^{-\alpha t} + A_2 t e^{-\alpha t} + K_1(t). \quad (4-19.8)$$

Передаточной функцией для $K(t)$ теперь служит

$$\varphi(\nu) = \frac{A_1}{\alpha + 2\pi i \nu} + \frac{A_2}{(\alpha + 2\pi i \nu)^2} + \varphi_1(\nu). \quad (4-19.9)$$

Следовательно, для больших ν :

$$\varphi_1(\nu) = \varphi(\nu) + \frac{A_1 i}{2\pi \nu} + \frac{A_2 + \alpha A_1}{(2\pi \nu)^2}. \quad (4-19.10)$$

Выберем теперь

$$\begin{aligned} A_1 &= -2\pi a_1, \\ A_2 &= -4\pi^2 a_2 - \alpha A_1. \end{aligned} \quad (4-19.11)$$

Тогда асимптотический закон (5) искажается поправочными членами, и $\varphi_1(\nu)$ стремится к нулю, как *третья степень* ν^{-1} . Разложение (7) функции $K_1(t)$ сходится теперь достаточно быстро. Степень α можно выбрать равной

$$\alpha = \frac{2\pi}{T_0}, \quad (4-19.12)$$

так как $e^{-2\pi}$ достаточно мало, чтобы можно было им пренебречь.

20. Интерполирование преобразования Фурье

Бесконечные пределы преобразования Фурье можно часто заменить конечными пределами, что иногда представляет большое преимущество. Даже если $f(t)$ определена в бесконечном интервале, мы можем вычесть из нее некоторую надлежаще выбранную функцию $f_0(t)$, которая при больших значениях t ведет себя асимптотически так же, как и $f(t)$. Мы, следовательно, заменяем первоначальную функцию $f(t)$ новой функцией $f_1(t) = f(t) - f_0(t)$, которая вне определенного интервала $t = \pm l$ практически равна нулю. Таким образом, мы получаем

$$F(\nu) = F_1(\nu) + F_0(\nu), \quad (4-20.1)$$

где $F_0(\nu)$ есть преобразование Фурье для функции $f_0(t)$, тогда как

$$F_1(\nu) = \int_{-l}^{+l} f_1(t) e^{-2\pi i \nu t} dt. \quad (4-20.2)$$

Эта новая функция имеет то замечательное свойство, что достаточно точно задать некоторое количество основных данных в качестве «узловых значений», и все остальные значения функции $F_1(\nu)$ получатся тогда интерполированием.

Мы принимаем, что $f_1(t)$ есть функция с ограниченной вариацией, определенная в интервале $(-l, l)$. Следовательно, $f_1(t)$ может быть разложена в сходящийся ряд Фурье

$$f_1(t) = \sum_{k=-\infty}^{+\infty} c_k e^{\frac{2\pi i k t}{2l}}, \quad (4-20.3)$$

где

$$c_k = \frac{1}{2l} \int_{-l}^{+l} f_1(t) e^{-\frac{\pi i k t}{l}} dt = \frac{1}{2l} F_1\left(\frac{k}{2l}\right). \quad (4-20.4)$$

Если внести бесконечный ряд (3) в формулу (2) и проинтегрировать почленно, то получим

$$F_1(\nu) = \frac{\sin 2\pi\nu l}{\pi} \sum_{k=-\infty}^{+\infty} F_1\left(\frac{k}{2l}\right) \frac{(-1)^k}{2l\nu - k}. \quad (4-20.5)$$

Это и есть интерполяционная (или экстраполяционная) формула. Она дает $F_1(\nu)$ для всех (вещественных или комплексных) значений ν в виде бесконечного сходящегося ряда, выраженного через основные ординаты

$$y_k = F_1\left(\frac{k}{2l}\right).$$

Если ν имеет вид $\nu = \frac{k}{2l} + \varepsilon$ и ε мало, то $F_1(\nu)$ будет определяться преимущественно ординатой y_k и непосредственно с ней смежными. Но если ε находится примерно посредине между двумя узловыми точками, то сходимость ряда (5) очень слаба. Мы можем ускорить сходимость с помощью следующего приема. Определим ряд ординат u_i по следующим рекуррентным формулам:

$$\left. \begin{aligned} u_0 &= y_0, \\ u_1 &= y_1 - u_0, & u_{-1} &= y_{-1} - u_0, \\ u_2 &= y_2 - u_1, & u_{-2} &= y_{-2} - u_{-1}, \\ &\dots & & \dots \end{aligned} \right\} \quad (4-20.6)$$

Тогда сумма (5) может быть заменена более быстро сходящейся суммой

$$F_1(\nu) = \frac{\sin 2\pi\nu l}{\pi} \left[-\frac{u_0}{2l\nu + 1} - \sum_{k=0}^{\infty} \frac{(-1)^k u_k}{(k - 2l\nu)(k - 2l\nu + 1)} + \right. \\ \left. + \sum_{k=1}^{\infty} \frac{(-1)^k u_{-k}}{(k + 2l\nu)(k + 2l\nu + 1)} \right]. \quad (4-20.7)$$

Здесь веса ординат убывают как квадраты их расстояний от центральной ординаты, и число членов, необходимых для удовлетворительной сходимости, становится сравнительно малым¹⁾).

21. Интерполяционный анализ фильтра

Мы видели в § 18, что частотное реагирование прибора связи есть преобразование Фурье его импульсной реакции. Ввиду того, что импульсная реакция $K(t)$ является не произвольной функцией от t , а исчезающей при всех отрицательных значениях t , то частотное реагирование есть отнюдь не произвольно выбираемая функция, а такая, которая должна удовлетворять некоторым вполне определенным аналитическим условиям, выраженным в соотношениях симметрии (18.17) и в интегральном условии (18.19). Но при конструировании электрических фильтров желательно получить $\varphi(\nu)$ с заранее известными свойствами; поэтому возникает вопрос, в какой мере фильтр может обладать определенными заданными свойствами и вместе с тем удовлетворять аналитическим условиям, обязательным вообще для отображений типа C .

При исследовании этой проблемы большую ценность представляет интерполяционная формула предыдущего параграфа. Как мы видели, импульсная реакция устойчивой цепи рассеивает энергию входа и, таким образом, практически становится равной нулю после определенного промежутка времени T_0 — «времени запоминания» цепи. Следовательно, функция $\varphi(\nu)$ имеет весьма определенный характер. Не только нижний предел интегрирования равен нулю, но и верхний предел есть конечное время T_0 :

$$\varphi(\nu) = \int_0^{T_0} K(t) e^{-2\pi i \nu t} dt. \quad (4-21.1)$$

Если теперь перейти от t к новой переменной t_1 путем преобразования

$$t = t_1 + \frac{1}{2} T_0 \quad (4-21.2)$$

и положить

$$K\left(t_1 + \frac{1}{2} T_0\right) = K_1(t_1), \quad (4-21.3)$$

то получим

$$\varphi(\nu) = e^{-\pi i \nu T_0} \varphi_1(\nu), \quad (4-21.4)$$

¹⁾ Ср. [10], стр. 441.

где

$$\varphi_1(\nu) = \int_{-\frac{T_0}{2}}^{\frac{T_0}{2}} K_1(t_1) e^{-2\pi i \nu t_1} dt_1. \quad (4-21.5)$$

Это выражение имеет рассмотренную уже раньше форму (20.2), в которой предел l заменен пределом $\frac{T_0}{2}$. Следовательно, может быть снова применена интерполяционная формула (20.5). Узловые значения суть

$$y_k = \varphi_1\left(\frac{k}{T_0}\right). \quad (4-21.6)$$

Они могут быть заданы в достаточной мере произвольно, однако с учетом условия вещественности

$$y_{-k} = y_k^* \quad (4-21.7)$$

и условия сходимости

$$\sum_{k=0}^{\infty} |y_k| = \text{конечному числу}. \quad (4-21.8)$$

Отметим, что в то время как даже в произвольно малом *непрерывном* интервале функция $\varphi(\nu)$ не может быть задана произвольно, $\varphi_1(\nu)$ практически может быть свободно задана в бесконечно большом числе равноотстоящих точек, соответствующих частотам

$$\nu = 0, \quad \nu_0 = \frac{1}{T_0}, \quad 2\nu_0, \quad 3\nu_0, \dots \quad (4-21.9)$$

Остается нерешенным вопрос, будет ли результат интерполяции функции $\varphi_1(\nu)$ *между* этими точками достаточно гладким. Но функция, порождаемая одной-единственной ординатой $y_k = 1$, дается формулой

$$\varphi_k(\nu) = \frac{\sin \pi (T_0 \nu - k)}{\pi (T_0 \nu - k)}. \quad (4-21.10)$$

Это выражение имеет характер ядра Дирихле (2.8). Ввиду вторичных максимумов этой функции концентрация в непосредственной окрестности точки $\nu = k\nu_0$ не очень велика. Поэтому интерполяция будет гладкой лишь в том случае, когда y_k изменяется гладко. Однако вблизи точки разрыва будут сказываться осложняющие колебания Гиббса (ср. § 9). Но мы видели, насколько успешно большие амплитуды этих колебаний могут быть снижены методом σ -сглаживания. Это сглаживание можно применить и в нашем случае. Оно состоит

в том, что мы берем локальное арифметическое среднее $\varphi(\nu)$ между $\nu + \frac{1}{T_0}$ и $\nu - \frac{1}{T_0}$.

В результате этой операции функция (10) заменяется функцией

$$\bar{\varphi}_k(\nu) = \frac{1}{2\pi} [\text{Si } \pi(T_0\nu - k + 1) - \text{Si } \pi(T_0\nu - k - 1)], \quad (4-21.11)$$

где

$$\text{Si } (\xi) = \int_0^\xi \frac{\sin x}{x} dx. \quad (4-21.12)$$

Функция Si называется «интегральным синусом»; это хорошо исследованная функция, для которой имеются таблицы. Найденная новая функция (10) удовлетворительно концентрируется в непосредственной окрестности центрального максимума.

Влияние операции сглаживания, примененной к равенству (5), состоит в том, что $K_1(t_1)$ заменяется функцией

$$\bar{K}_1(t_1) = K_1(t_1) \frac{\sin\left(\frac{2\pi t_1}{T_0}\right)}{2\pi t_1} T_0. \quad (4-21.13)$$

Сама функция $K_1(t_1)$ может быть выражена через заданные ординаты y_k согласно формуле (20.3):

$$K_1(t_1) = \frac{1}{T_0} \sum_{k=-\infty}^{+\infty} y_k e^{-\frac{2\pi i k t_1}{T_0}}. \quad (4-21.14)$$

Одной из классических задач в теории фильтров является конструирование «идеального фильтра» нижних частот. Этот фильтр должен срезать все частоты выше определенной граничной частоты ν_c , но пропускать все частоты ниже ν_c без какого бы то ни было искажения амплитуды или фазы. Непосредственное решение этой задачи может быть получено методом интерполяции и последующего исключения колебаний Гиббса путем σ -сглаживания. Импульсная реакция, соответствующая такому фильтру, получается по формуле

$$\bar{K}_1(t_1) = \begin{cases} \frac{\sin(2\pi\nu_c t_1) \cos \pi t_1 / T_0}{2\pi t_1} & \left(|t_1| \leq \frac{T_0}{2}\right), \\ 0 & \left(|t_1| > \frac{T_0}{2}\right). \end{cases} \quad (4-21.15)$$

До сих пор мы занимались функцией $\varphi_1(\nu)$, а не первоначальной функцией $\varphi(\nu)$. Соотношение между этими двумя функциями дается формулой (4). Значение функции $\varphi(\nu)$, по определению, задается

следующими формулами:

в х о д:
$$f(t) = e^{2\pi i \nu t},$$

в ы х о д:
$$g(t) = \varphi(\nu) e^{2\pi i \nu t} = \varphi_1(\nu) e^{2\pi i \nu \left(t - \frac{T_0}{2}\right)}.$$

Отсюда видно, что выход, рассчитанный на основе функции $\varphi_1(\nu)$, отстает от входа на постоянный отрезок времени $\frac{T_0}{2}$. Следовательно, свобода, которой мы располагаем при конструировании фильтров с заданными характеристиками, ограничивается тем, что существует неизбежное *постоянное отставание во времени*, которое равно половине времени запоминания, связанного с использованием фильтра. Во многих задачах связи это отставание во времени не имеет последствий. Но если мы не можем допустить этого отставания и должны свести T_0 до незначительной величины, то постепенно теряется точность в реализации характеристики фильтра, так как основная частота $\nu_0 = \frac{1}{T_0}$ велика, и точки, для которых может быть задана функция $\varphi(\nu)$, слишком сильно расходятся друг от друга. В этом случае мы упускаем из виду важные характеристики исследуемой кривой.

22. Разыскание скрытых периодичностей

В некоторых метеорологических и астрономических задачах, в анализе приливов и во всех положениях, при которых предполагаются скрытые периодичности, мы встречаемся со следующей математической задачей. Известно, что некоторая функция разлагается на строго периодические составляющие, хотя эти составляющие не находятся в гармоническом отношении друг с другом*). Требуется найти неизвестные частоты и амплитуды каждой из составляющих. В нашем распоряжении имеется большое число наблюдений, сделанных через равные промежутки времени. Выводы должны быть получены на основании информации, которую доставили нам эти наблюдения.

Предположим, что общее число наблюдений есть нечетное число $2N+1$. Кроме того, примем за начальный момент отсчета времени $t=0$ момент, соответствующий средней точке наших наблюдений. Пусть наблюдения производились через интервалы времени τ . Перейдем от первоначальной переменной t_1 к новой переменной

$$t = \frac{1}{\tau} t_1. \quad (4-22.1)$$

Тогда наблюдения будут относиться к моментам времени

$$t_k = 0, \pm 1, \pm 2, \dots, \pm N. \quad (4-22.2)$$

*) То есть отношение их периодов не равно рациональному числу.

Для отсчетов введем обозначение:

$$f_k = f(t_k). \quad (4-22.3)$$

Функция $f(t)$, вообще говоря, имеет следующую форму:

$$f(t) = \sum_{\alpha=1}^j (A_\alpha \cos \theta_\alpha t + B_\alpha \sin \theta_\alpha t). \quad (4-22.4)$$

Число периодических компонент, обозначенное здесь через j , обычно заранее неизвестно. Точно так же мы ничего не можем сказать о величинах угловых частот $\omega_\alpha = \frac{\theta_\alpha}{\tau}$. Однако мы заранее знаем, что в новой переменной t угловые частоты θ_α должны ограничиваться неравенством

$$\theta_\alpha < \pi, \quad (4-22.5)$$

так как, если θ_α превышает этот предел, то две частоты $\pi + \beta$ и $\pi - \beta$ не могут быть различимы. Мы положим

$$\theta_\alpha = \frac{\pi}{N} p_\alpha, \quad (4-22.6)$$

и пусть p_α принимает значения между 0 и N .

Отделим синусы от косинусов, образовав суммы и разности

$$\begin{aligned} f(t) + f(-t) &= 2 \sum_{\alpha=1}^j A_\alpha \cos \theta_\alpha t, \\ f(t) - f(-t) &= 2 \sum_{\alpha=1}^j B_\alpha \sin \theta_\alpha t. \end{aligned} \quad (4-22.7)$$

Наши ординаты разделятся соответственно на две группы:

$$\begin{aligned} u_k &= f_k + f_{-k}, \\ v_k &= f_k - f_{-k} \end{aligned} \quad (k=0, 1, 2, \dots, N). \quad (4-22.8)$$

Мы воспользуемся методом преобразования Фурье [ср. (17. 8)], приспособленным, однако, для суммирования вместо интегрирования. Первоначальный ряд данных u_k мы преобразуем в новый ряд $N+1$ амплитуд a_k при косинусах, а данные v_k — в ряд $N-1$ амплитуд b_k при синусах. Эти амплитуды получаются умножением исходных данных на заранее составленную матрицу, содержащую косинусы и синусы углов, кратных $\frac{\pi}{N}$:

$$\left. \begin{aligned} a_k &= \sum_{\alpha=0}^N u_\alpha \cos \frac{\pi}{N} \alpha k, \\ b_k &= \sum_{\alpha=1}^{N-1} v_\alpha \sin \frac{\pi}{N} \alpha k. \end{aligned} \right\} \quad (4-22.9)$$

Штрих в символе $\sum'$ отмечает тот факт, что первое и последнее данные значения функции u_0 и u_N входят в сумму с множителем $\frac{1}{2}$.

Полученные коэффициенты a_k и b_k можно представить себе как «линейный спектр», соответствующий целым значениям p_α непрерывного параметра p . Весь дальнейший анализ будет применять эти две новые последовательности коэффициентов.

Рассмотрим сначала случай, когда заданные частоты θ_α таковы, что все p_α в формуле (6) оказываются *целыми числами*. Тогда амплитуды (9) непосредственно дают решение нашей задачи. Большинство коэффициентов a_k и b_k будет равно нулю. Если некоторый коэффициент a_k или b_k не равен нулю, то это указывает, что наши исходные данные содержат частоту

$$\theta = \frac{\pi}{N} k. \quad (4-22.10)$$

Соответствующая амплитуда при косинусе будет

$$A = \frac{1}{N} a_k, \quad (4-22.11)$$

а соответствующая амплитуда при синусе —

$$B = \frac{1}{N} b_k. \quad (4-22.12)$$

В общем случае точное значение p , соответствующее некоторой частоте θ_α , будет лежать между двумя целыми числами m и $m+1$, которые распознаются по тому факту, что регулярное чередование знаков $\pm$ амплитуд a_k и b_k прерывается следованием $++$ или $--$. Теперь мы полагаем

$$p_\alpha = m + \varepsilon \quad (4-22.13)$$

и получаем ε при помощи интерполирования. Для этой цели пользуемся методом «вторых сумм» [ср. (3-5.28)], описанным в III, 5. В этом методе взаимное влияние «пиковых» значений значительно уменьшается. После составления отношения q двух соседних вторых сумм [ср. (3-5.30)] ε получается по формуле (3-5.31). Превосходной проверкой точности наших наблюдений служит тот факт, что положение «пиковых» значений амплитуд как a_k , так и b_k должно быть одним и тем же. Следовательно, мы имеем два независимых определения частот θ_α : одно с помощью a_k и другое с помощью b_k .

После определения ε мы получаем частоту θ_α , используя соотношения

$$\theta_\alpha = \frac{\pi}{N} (m + \varepsilon). \quad (4-22.14)$$

Кроме того, мы можем выразить коэффициенты A_α , B_α , соответствующие этой частоте, через амплитуды ряда Фурье a_m и b_m :

$$A_\alpha = \frac{a_m}{N} \frac{\pi \epsilon}{\sin \pi \epsilon}, \quad B_\alpha = \frac{b_m}{N} \frac{\pi \epsilon}{\sin \pi \epsilon}. \quad (4-22.15)$$

Описанный метод выделения периодических составляющих весьма удовлетворителен, если общее число наблюдений $2N+1$ достаточно велико для того, чтобы изолировать два соседних «пиковых» значения. Необходимо, чтобы два последовательных значения p_α в соотношении (6) были отделены друг от друга по меньшей мере четырьмя единицами. Если они расположены ближе друг к другу, то их взаимное влияние возрастает; в этом случае трудно выделить их надлежащим образом.

Если мы не располагаем достаточным числом наблюдений, а пиковые значения очень близки друг к другу, то целесообразно избрать совершенно другой подход. Как мы видели ранее (ср. § 21), фокусирующее действие функции (2.8) можно значительно усилить методом σ -сглаживания. В нашей задаче этот метод сводится к видоизменению исходных данных u_k и v_k , которые входят в суммы Фурье (9). Прежде чем образовать эти суммы, мы изменяем исходные данные, умножая их на надлежаще выбранные весовые множители согласно следующим формулам:

$$\bar{u}_k = u_k \sigma_k, \quad \bar{v}_k = v_k \sigma_k, \quad (4-22.16)$$

в которых

$$\sigma_k = \frac{\sin \frac{k\pi}{N}}{\frac{k\pi}{N}}. \quad (4-22.17)$$

Коэффициенты a_k , b_k в формуле (9) вычисляются теперь с помощью этих новых величин $\bar{u}_\alpha$, $\bar{v}_\alpha$.

В результате этих изменений функция $\frac{\sin \pi x}{\pi x}$ заменяется функцией

$$S(x) = \frac{1}{2\pi} [\text{Si}(x + \pi) - \text{Si}(x - \pi)]. \quad (4-22.18)$$

Вторичные пики этой функции составляют 4,8; 2,0; 1,1%, ... от центрального максимума сравнительно с 21,7; 12,8; 9,1%, ... для первоначальной функции (см. рис. 34 в § 7). Таким образом, новая функция обладает более сильным фокусирующим действием, а ее новые максимумы практически независимы. В то же время пики теперь несколько шире, чем они были раньше.

Это расширение в расположении ординат имеет то благоприятное влияние, что мы можем провести параболу второго порядка через максимальную амплитуду и два соседних с ней максимума, получив

с помощью этой параболической интерполяции положение и величину истинного максимума. Возьмем максимальную амплитуду a_m и левый и правый соседние максимумы a_{m-1} , a_{m+1} . Величина ε , входящая в формулы (13) и (14), теперь определится следующим образом:

$$\varepsilon = \frac{1}{2} \frac{a_{m+1} - a_{m-1}}{2a_m - (a_{m+1} + a_{m-1})}, \quad (4-22.19)$$

а максимальная ордината a_μ — по формуле

$$a_\mu = a_m + \frac{\varepsilon}{4} (a_{m+1} - a_{m-1}). \quad (4-22.20)$$

Тогда

$$A_\alpha = \frac{a_\mu}{N} \cdot 1,6963 \quad (4-22.21)$$

(численный множитель есть величина, обратная значению $S(0)$). Такое же вычисление можно выполнить для амплитуды B_α при синусе, заменив a_k через b_k ¹⁾.

23. Выделение показательных функций

В предыдущем параграфе был разобран вопрос об анализе функции, состоящей из периодических составляющих. При измерениях радиоактивного распада возникает аналогичная задача, но здесь роль периодических функций играют показательные функции. Пусть задана функция следующего вида:

$$f(x) = A_1 e^{-\lambda_1 x} + A_2 e^{-\lambda_2 x} + \dots + A_m e^{-\lambda_m x}. \quad (4-23.1)$$

Требуется найти постоянные распада λ_i и активности A_i . Внешне эта задача весьма сходна с предыдущей: частота ω_i заменена мнимой частотой $i\omega_i$. Однако на самом деле эти две задачи очень далеки друг от друга, так как показательные функции не обладают теми замечательными свойствами ортогональности, которые присущи периодическим функциям. Ввиду этого недостаточно лишь разобрать чисто математическое решение задачи, надо также рассмотреть и ее численные аспекты. Мы сначала займемся теоретической стороной, а затем перейдем к численным следствиям.

Как известно, решение обыкновенного линейного дифференциального уравнения с постоянными коэффициентами может быть представлено в виде линейной комбинации показательных функций. Таким образом, решение имеет вид (1). Это же справедливо и для разностного линейного уравнения с постоянными коэффициентами, в котором

¹⁾ Численные схемы, разобранные в этом параграфе, гораздо точнее традиционных методов «периодограмм», применение которых не дает возможности распознать фундаментальную связь между задачей скрытых периодичностей и теорией преобразования Фурье.

элемент. Наконец, точно так же как это делается при обыкновенном умножении чисел, составляются суммы каждого столбца:

$$\begin{array}{ccccccc}
 f_1 y_0, & f_2 y_0, & \dots, & f_{m-1} y_0, & y_0, & & \\
 f_2 y_1, & f_3 y_1, & \dots, & y_1, & & & \\
 \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
 f_{m-1} y_{m-2}, & y_{m-2}, & & & & & \\
 y_{m-1} & & & & & & \\
 \hline
 \text{Суммы: } g_0, & g_1, & \dots, & g_{m-2}, & g_{m-1} & &
 \end{array} \quad (4-23.11)$$

Это дает полином

$$G_{m-1}(p) = g_0 + g_1 p + g_2 p^2 + \dots + g_{m-1} p^{m-1}. \quad (4-23.12)$$

Теперь неизвестные A_i нашей задачи (8) получаются с помощью простого процесса подстановки. Мы подставляем $p = p_i$ в $G_{m-1}(p)$ и в производную от функции $F_m(p)$. Отношение этих двух чисел дает A_i :

$$A_i = \frac{G_{m-1}(p_i)}{F'_m(p_i)} \quad (4-23.13)$$

Другой метод решения системы (8) состоит в построении матрицы, обратной матрице линейной системы (8). Если элементы обратной матрицы обозначить через q_{ik} , то

$$q_{ik} = \frac{F_{k-1}(p_i)}{F'_m(p_i)} \quad (i, k = 1, 2, \dots, m), \quad (4-23.14)$$

где полином $F_{k-1}(p)$ образуется с помощью *последних* элементов основного полинома $F_m(p)$, например

$$F_2(p) = f_{m-2} + f_{m-1} p + p^2.$$

Последовательные числители элементов q_{ik} (для фиксированного i) можно получить также с помощью алгебраического деления (ср. I, 8). Если $F_m(p)$ разделить на двучлен $p - p_i$, то последовательные коэффициенты при степенях p в частном будут

$$F_0(p_i), F_1(p_i), \dots, F_{m-1}(p_i).$$

Эта простая схема решения задачи выделения показательных функций не дает еще представления о тех огромных практических трудностях, которые возникают, когда ее пытаются применить к физическим задачам. Эти трудности обуславливаются тем фактом, что решить уравнения (7) относительно коэффициентов c_i удастся лишь тогда, когда известные величины u задаются с *исключительной точностью*. Если требуется выделить четыре или пять показательных функций, то соответствующая линейная система (7) становится настолько «косоугольной», что для ее решения требуется задать величины u ,

с точностью до шести-восьми значащих цифр. Такая точность совершенно нереалистична, если учесть фактическую точность измерений распада. Даже выделение трех показательных функций может встретиться с неодолимыми трудностями. Нижеследующий пример хорошо иллюстрирует те поразительные каверзы, которые может сделать неортогональный характер экспоненциальных функций.

Следующие 24 наблюдения распада были сделаны с интервалом в три минуты, т. е. в 0,05 часа, если считать час единицей времени в нашей задаче; следовательно, $\Delta x = 0,05$ при начальном моменте времени $x = 0$:

k	y_k	k	y_k	k	y_k	k	y_k
1	2,51	7	0,77	13	0,27	19	0,11
2	2,04	8	0,64	14	0,23	20	0,10
3	1,67	9	0,53	15	0,20	21	0,09
4	1,37	10	0,45	16	0,17	22	0,08
5	1,12	11	0,38	17	0,15	23	0,07
6	0,93	12	0,32	18	0,13	24	0,06

Наблюдения считаются точными до половины второго десятичного знака.

Мы не знаем, сколько показательных функций содержат наши данные. Сначала мы попытаемся выделить *три* функции. Поэтому разобьем наши данные на шесть групп, беря в каждой группе сумму четырех последовательных ординат с k от 1 до 4, от 5 до 8 и т. д. Опуская в этих суммах запятые, которые для наших целей не имеют значения, получим следующие новые данные: 759, 346, 168, 87, 49, 30.

Уравнения для определения коэффициентов c_i в этом случае запишутся так:

$$\left. \begin{aligned} 759 c_0 + 346 c_1 + 168 c_2 + 87 &= 0, \\ 346 c_0 + 168 c_1 + 87 c_2 + 49 &= 0, \\ 168 c_0 + 87 c_1 + 49 c_2 + 30 &= 0. \end{aligned} \right\}$$

Деля на множитель при c_0 , получим

$$\left. \begin{aligned} c_0 + 0,4559 c_1 + 0,2213 c_2 + 0,1146 &= 0, \\ c_0 + 0,4855 c_1 + 0,2514 c_2 + 0,1416 &= 0, \\ c_0 + 0,5179 c_1 + 0,2917 c_2 + 0,1786 &= 0. \end{aligned} \right\}$$

Исключая отсюда c_0 , придем к следующим двум уравнениям для определения c_1 и c_2 :

$$\left. \begin{aligned} 0,0296 c_1 + 0,0301 c_2 + 0,0270 &= 0, \\ 0,0324 c_1 + 0,0402 c_2 + 0,0370 &= 0. \end{aligned} \right\} \tag{4-23.15}$$

Умножив второе уравнение на 0,75, получим

$$0,0242 c_1 + 0,0302 c_2 + 0,0277 = 0.$$

Но точность наших наблюдений такова, что мы не можем в коэффициенте ручаться больше, чем за единицу второго знака после запятой. А это показывает, что система уравнений (15) в пределах ошибок наших наблюдений является избыточной. Она никоим образом не может служить для независимого определения коэффициентов c_1 и c_2 . Отсюда следует, что наши измерения могут быть описаны с помощью только двух показательных функций. Поэтому мы разбиваем теперь наши данные на четыре группы по шести ординат в каждой и составляем суммы внутри каждой группы.

Это дает новые данные: 964, 309, 115, 51, и мы получаем следующие два уравнения:

$$\begin{aligned} 964 c_0 + 309 c_1 + 115 &= 0, \\ 309 c_0 + 115 c_1 + 51 &= 0, \end{aligned}$$

имеющие решение

$$c_0 = 0,1640, \quad c_1 = -0,8839.$$

Это приводит к квадратному уравнению

$$\xi^2 - 0,8839 \xi + 0,1640 = 0,$$

корни которого

$$\xi_1 = 0,619, \quad \xi_2 = 0,265.$$

Значение h в формуле (6) теперь равно $6 \cdot 0,05 = 0,3$; таким образом, показатели λ_1 и λ_2 равны

$$\lambda_1 = 4,45, \quad \lambda_2 = 1,58.$$

Теперь мы переходим к определению двух амплитуд A_1 и A_2 . В принципе для этой цели достаточно двух наблюдений. Мы используем, однако, излишек данных для проверки постоянства амплитуд A_i .

Линейная система (8) может быть решена следующим образом. Подберем линейную комбинацию m равноотстоящих данных, при которой веса (коэффициенты) всех A_α станут равными нулю, за исключением веса при одном A_i . Если теперь применить эту линейную комбинацию последовательно ко всем данным таблицы, то мы придем к выделению одной показательной функции $A_i e^{-\lambda_i x}$. Например, в нашей задаче мы можем скомбинировать данные с k от 1 до 5 для уничтожения A_2 . Но затем мы можем взять такую же линейную комбинацию данных с k от 2 до 6, затем от 3 до 7 и т. д. Таким образом, мы последовательно получаем разные значения для величины $A_i e^{-\lambda_i x}$. Если теперь умножить их на $e^{\lambda_i x}$, то мы должны получить одну и ту же величину, если не считать неизбежного разброса, обусловленного неточностью наших данных. Если расхождения имеют

линейный характер, то методом наименьших квадратов находим соответствующий линейный двучлен

$$A + Bx = A(1 + \beta x), \quad \beta = \frac{B}{A}.$$

Ввиду малости β можно принять, что $A + Bx = Ae^{\beta x}$, и использовать полученную величину β в качестве поправки к найденному ранее значению показателя λ_1 :

$$Ae^{\beta x}e^{-\lambda_1 x} = Ae^{-(\lambda_1 - \beta)x}.$$

Проведя этот процесс в нашем примере, мы найдем, что расхождения не обнаруживают линейного характера и что имеется лишь естественный разброс результатов. Например, первые 12 последовательных значений для определения A_1 с точностью до общего множителя пропорциональности оказываются равными

$$887, 870, 872, 842, 872, 867, 861, 869, 862, 908, 890, 914.$$

Мы остановились на значении 914, так как увеличение, вызываемое умножением на $e^{\lambda_1 x}$ здесь уже составляет 14,4. Взяв арифметическое среднее этих значений, получаем 876,1, что приводит к значению

$$A_1 = 2,202.$$

Аналогичное положение получается и для A_2 . Здесь также последовательный ряд 16 значений для A_2 не указывает на существование линейного остатка. Арифметическое среднее дает

$$A_2 = 0,305.$$

Таким образом, окончательный результат имеет вид

$$f(x) = 2,202e^{-4,45x} + 0,305e^{-1,58x}. \tag{4-23.16}$$

Эта формула дает математическое представление наших данных с помощью двух показательных функций. В виде контроля мы определяем значения этой функции для всех 24 данных точек $x = 0; 0,05; 0,1; 0,15; \dots; 1, 15$.

<i>k</i>	<i>y_k</i>	<i>k</i>	<i>y_k</i>	<i>k</i>	<i>y_k</i>	<i>k</i>	<i>y_k</i>
1	2,507	7	0,769	13	0,270	19	0,114
2	2,044	8	0,639	14	0,230	20	0,100
3	1,672	9	0,533	15	0,197	21	0,088
4	1,370	10	0,447	16	0,173	22	0,079
5	1,126	11	0,376	17	0,148	23	0,070
6	0,929	12	0,318	18	0,130	24	0,063

Сравнение этой таблицы с первоначальной таблицей (стр. 284) наших данных показывает, что расхождение нигде не превышает 0,005, за

исключением одного случая $k=5$, где ошибка достигает величины 0,006. Можно охарактеризовать «среднее отклонение», взяв корень квадратный из суммы квадратов всех отдельных отклонений, деленный на 24. Мы получаем при этом лишь 0,0026, что хорошо укладывается в пределы ошибок наших данных. Следовательно, можно считать, что закон (16) вполне удовлетворительно представляет наши данные.

На самом же деле, мы потерпели жестокую неудачу в нашей работе, ибо предложенные данные были получены на основе закона

$$f(x) = 0,0951 e^{-x} + 0,8607 e^{-3x} + 1,5576 e^{-5x}. \quad (4-23.17)$$

Мы не только потеряли относительно слабо представленную третью показательную функцию, но присутствие этой компоненты оказало искажающее влияние на другие две показательные функции, уменьшив показатель 3 до 1,58 и показатель 5 до 4,45. Кроме того, отношение амплитуд, составляющее приблизительно 1:2, искажено до 1:7. Наш результат, следовательно, совершенно неудовлетворителен. И все же наше решение «численно эквивалентно» правильному решению, так как дает вполне удовлетворительное совпадение с данными в пределах ошибок измерений. Не следует надеяться, что какой-нибудь иной математический прием мог бы привести к лучшим результатам, ибо трудность заключается не в способе вычисления, а в исключительной чувствительности показательных функций и амплитуд к весьма малым изменениям данных. От такого влияния не могут нас избавить никакие наименьшие квадраты или другие статистические приемы. Единственным средством было бы увеличение точности измерений до таких пределов, которые далеко превышают возможности современных измерительных приборов.

При условиях ограниченной точности измерений мы сделаем более скромную попытку. Допустим, что мы имеем некоторую предварительную информацию относительно показателей λ_i ; точнее, мы знаем число показательных компонент и приблизительное значение каждого показателя. Наша задача теперь сводится к определению амплитуд A_i , а также поправок к заданным λ_i .

Мы снова будем действовать подобно тому, как мы это делали в последней фазе нашего предыдущего решения. На основе заданных λ_i мы строим такую линейную комбинацию данных, которая выделяет одну показательную функцию $A_i e^{-\lambda_i x}$. Применяя эту же линейную комбинацию к ряду данных, мы получаем последовательность, скажем, $2n$ равноотстоящих значений функции. Делим сумму первых n значений на сумму вторых n значений; это дает

$$e^{\lambda_i n \Delta x},$$

откуда можно определить λ_i . Мы получили, таким образом, одно уточненное значение λ_i ; воспользуемся теперь этим новым значением λ

вместе с остальными заданными λ для того, чтобы провести исправление другого из показателей λ_i . Продолжаем процесс таким же образом, принимая во внимание на каждом новом этапе уточненные значения показателей; получив уточненные значения всех показателей, определим амплитуды A_i таким же образом, как это мы делали раньше.

Применим этот прием к предыдущему примеру. Пусть заданные значения показателей будут

$$\lambda_1 = 1,2 \text{ (вместо 1)}, \lambda_2 = 2,7 \text{ (вместо 3)}, \lambda_3 = 6 \text{ (вместо 5)}.$$

В результате наших вычислений получаем:

$$f(x) = 0,041 e^{-0,50x} + 0,79 e^{-2,73x} + 1,68 e^{-4,96x}. \quad (4-23.18)$$

Сравнительно с точным выражением (17) результат снова разочаровывающий. И опять мы получаем вполне удовлетворительное согласие с заданными величинами в пределах точности наших измерений. Этот пример показывают, что мы не можем ожидать удовлетворительных результатов при решении задачи выделения показательных функций даже в случае, когда мы заранее знаем приблизительные значения постоянных распада и нам надо лишь уточнить эти значения. С другой стороны, картина была бы совершенно иная, если бы наши данные были в десять раз более точными *).

24. Преобразование Лапласа

К теории преобразования Фурье тесно примыкает несколько отличное от него преобразование, которое играет исключительную роль во многих задачах математической физики и оказывается особенно полезным в задачах анализа цепей. Это — так называемое «преобразование Лапласа», определяемое формулой

$$\mathcal{L}(z) = \int_0^{\infty} e^{-xz} f(x) dx. \quad (4-24.1)$$

При мнимых значениях переменной z формула (1) принимает вид (17.16), т. е. определяет преобразование Фурье для функции $f(x)$. Теперь эта функция $f(x)$ должна быть задана лишь в интервале $(0, \infty)$, так как для отрицательных значений x мы принимаем $f(x) = 0$. Однако новый нижний предел интегрирования 0 вместо $-\infty$ оказывает глубокое влияние на аналитическую природу функции $\mathcal{L}(z)$. В то время как прежняя функция $F(y)$ (ср. 17.16) была определена лишь для *вещественных* значений переменной y , новое преобразование представляет собой функцию комплексного перемен-

*) Дополнения к изложенному в этом параграфе материалу читатель найдет в заметке [11].

ного z на всей комплексной полуплоскости $R(z) > 0$, если только $f(x)$ есть абсолютно интегрируемая функция от x в бесконечном интервале $(0, \infty)$. Если $f(x)$ недостаточно быстро убывает, чтобы быть абсолютно интегрируемой, может все же случиться, что $f(x)e^{-\varepsilon x}$ уже интегрируема при произвольно малом ε . Тогда мы можем положить

$$z = \varepsilon + z_1, \quad f_1(x) = f(x)e^{-\varepsilon x} \quad (4-24.2)$$

и рассматривать преобразование Лапласа для функции $f_1(x)$. Мы теперь тогда для нашей первоначальной переменной z узкую полосу шириною в ε , расположенную около мнимой оси. Так как, однако, ε может быть сделано произвольно малым, то аналитический характер функции $\mathcal{L}(z)$ повсюду справа от мнимой оси все же обеспечен. Например, при $f(x) = 1$ получаем

$$\mathcal{L}(z) = \int_0^{\infty} e^{-zx} dx = \frac{1}{z}. \quad (4-24.3)$$

Эта функция от z , первоначально определенная лишь для $R(z) > 0$, может теперь быть распространена путем аналитического продолжения на всю комплексную плоскость, за исключением лишь особой точки $z = 0$.

Изменяя одновременно знаки переменных x и z , мы можем определить преобразование Лапласа $\mathcal{L}_1(z)$ для отрицательной полуплоскости:

$$\mathcal{L}_1(z) = \int_{-\infty}^0 e^{-xz} f_1(x) dx \quad (4-24.4)$$

и затем объединить $\mathcal{L}(z) + \mathcal{L}_1(z)$ в одну функцию. Общей границей обоих определений служит мнимая ось, где теперь имеет место равенство

$$\mathcal{L}(i\omega) + \mathcal{L}_1(i\omega) = \int_{-\infty}^{+\infty} e^{-i\omega x} f(x) dx. \quad (4-24.5)$$

Это соотношение показывает, что преобразование Фурье абсолютно интегрируемой функции в интервале $(-\infty, +\infty)$ можно рассматривать как сумму двух преобразований Лапласа, взятых вдоль мнимой оси, из которых одно есть функция, аналитическая в правой полуплоскости, а другое — функция, аналитическая в левой полуплоскости.

Между функцией $f(x)$ и преобразованием Лапласа $\mathcal{L}(z)$ не существует точечного соответствия. Значение функции $\mathcal{L}(z)$ в любой точке зависит от всей совокупности значений $f(x)$. Имеется, однако, одно исключение. Предположим, что $f(x)$ — аналитическая функция в окрестности $x = 0$. Дадим теперь переменной z большое вещественное положительное значение. Тогда в силу того, что

функция e^{-zx} весьма мала для любого x , существенно отличного от нуля, область интегрирования сводится к весьма малой окрестности $x=0$. При этих условиях мы получаем

$$\int_0^{\infty} f(x) e^{-xz} dx = f(0) \int_0^{\infty} e^{-xz} dx = \frac{f(0)}{z}. \quad (4-24.6)$$

В более общем случае, разлагая $f(x)$ в окрестности $x=0$ в сходящийся ряд Тейлора

$$f(x) = c_0 + c_1 x + c_2 x^2 + \dots \quad (4-24.7)$$

и интегрируя почленно, получаем для соответствующего преобразования Лапласа при больших значениях z (в предположении, что вещественная часть переменной z положительна):

$$\mathcal{L}(z) = \frac{c_0}{z} + \frac{c_1}{z^2} + \frac{2!c_2}{z^3} + \frac{3!c_3}{z^4} + \dots \quad (4-24.8)$$

Эта формула осуществляет точечное соответствие между функцией $f(x)$ в бесконечно малой окрестности точки $x=0$ и преобразованием Лапласа в бесконечно малой окрестности точки $z=\infty$. Однако ряд (8) не обязательно сходится при каждом значении z .

25. Анализ цепей и преобразование Лапласа

Применение преобразования Лапласа к уравнениям электрических цепей открыло новую перспективу, которую предвосхитили операционные методы Хевисайда. Остроумная интуиция Хевисайда получила свое оправдание, когда было обнаружено, что преобразования Лапласа функций входа и выхода электрических цепей автоматически удовлетворяют алгебраическим уравнениям, к которым Хевисайд пришел несколько менее строгим путем. Система обыкновенных дифференциальных уравнений с постоянными коэффициентами переходит в систему простых линейных алгебраических уравнений, связывающих преобразования Лапласа первоначальных неизвестных. Если задана блок-схема электрической цепи, то соотношение входа — выхода электрической цепи может быть получено путем решения системы линейных уравнений. Это решение дает зависимость входа — выхода в форме отношения двух алгебраических полиномов: $p(z)$ и $q(z)$. Основной величиной здесь является импульсная реакция цепи. Преобразование Лапласа единичного импульса равняется 1. Преобразование Лапласа выходной функции, т. е. импульсной реакции:

$$\mathcal{L}(z) = \int_0^{\infty} e^{-zt} K(t) dt. \quad (4-25.1)$$

Именно эта функция $\mathcal{L}(z)$ непосредственно может быть выведена из заданной блок-схемы цепи и представлена рациональной дробью

$$\mathcal{L}(z) = \frac{p(z)}{q(z)}, \quad (4-25.2)$$

где степень $q(z)$ по крайней мере на единицу меньше, чем степень $p(z)$.

Первая физическая величина, которую мы теперь непосредственно получим — это *частотная реакция* цепи. Частотная реакция $\varphi(\omega)$, выраженная через угловую частоту [ср. (17.19)], получается из $\mathcal{L}(z)$ в результате подстановки $(i\omega)$ вместо z :

$$\varphi(\omega) = \mathcal{L}(i\omega). \quad (4-25.3)$$

Таким образом, частотная реакция может быть получена без интегрирования путем использования только заданных сведений об элементах цепи и их коммутации. С другой стороны, импульсная реакция получается как результат преобразования Фурье [ср. (18.10)]:

$$K(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \mathcal{L}(i\omega) e^{i\omega t} d\omega. \quad (4-25.4)$$

Это интегрирование, однако, — как было отмечено Хевисайдом, — нет нужды проводить, так как можно разложить рациональную дробь (2) на простые дроби

$$\frac{p(z)}{q(z)} = \sum_{i=1}^n \frac{a_i}{z - \lambda_i}. \quad (4-25.5)$$

Здесь λ_i суть корни знаменателя, а коэффициенты a_i определяются по формулам*)

$$a_i = \frac{p(\lambda_i)}{q'(\lambda_i)}.$$

После элементарного интегрирования

$$\int_0^{\infty} e^{\lambda_i t - zt} dt = \frac{1}{z - \lambda_i} \quad (4-25.6)$$

мы получаем:

$$K(t) = \sum_{i=1}^n a_i e^{\lambda_i t}. \quad (4-25.7)$$

Это — знаменитая «теорема разложения» Хевисайда (в слегка видоизмененной форме, так как Хевисайд интересовался реакцией на ступенчатую функцию, которая представляет собой интеграл от импульсной реакции).

*) В предположении, что знаменатель $q(z)$ не имеет кратных корней.

26. Обращение преобразования Лапласа

Теорема разложения Хевисайда и решает задачу нахождения неустановившегося режима в электрических цепях; однако на практике часто предпочитают другие методы, в которых не возникает необходимости нахождения комплексных корней алгебраического уравнения. Действительно, в случае сложных блок-схем, встречающихся в теории сервомеханизмов для управляемых ракет, даже фактическое построение полиномов $p(z)$, $q(z)$ может натолкнуться на непреодолимые практические трудности, хотя получение численного значения функции $\mathcal{L}(z)$ для любого заданного значения z может оказаться не очень трудным делом. Поэтому можно поставить следующий вопрос: задано преобразование Лапласа некоторой функции $f(x)$ («исходной функции») в определенных, надлежаще выбранных, точках комплексной плоскости; найти первоначальную (исходную) функцию $f(t)$ ¹⁾. Так как интервал переменной t бесконечен, то масштаб, которым измеряется t , в принципе произволен. Фактически же надлежащая калибровка переменной t чрезвычайно важна. Первоначальный масштаб, которым измерено t , может оказаться совершенно неподходящим для нашей задачи. В задачах теории электрических цепей функция $K(t)$ имеет значение импульсной реакции цепи. Ввиду практически конечного «времени запоминания» T_0 сети функция $K(t)$ представляет интерес только вплоть до определенного T_0 , которое, однако, вообще говоря, может быть заранее неизвестно. Если единица времени выбрана слишком большой, то существенная часть функции $K(t)$ может оказаться втиснутой в очень малый промежуток около $t=0$. Если, с другой стороны, единица времени выбрана слишком малой, существенная часть $K(t)$ может быть слишком растянута. В обоих случаях может пострадать сходимость рядов, которые нам предстоит изучить в ближайших параграфах.

Время запоминания хорошо было бы выбрать в качестве естественной единицы времени. Однако, и не зная этой величины, все же мы можем ввести целесообразное, хотя и не всегда надежное, нормирование масштаба времени. Даже не произведя подробного исследования распределения корней знаменателя $q(z)$, можно составить некоторое суждение относительно их общего расположения. Пусть коэффициент при z^n в $q(z)$ приведен к 1. Тогда коэффициент q_{n-1} при степени z^{n-1} имеет следующее значение: он равен сумме всех корней, взятой с обратным знаком. Следовательно, формула

$$-\frac{q_{n-1}}{n} = \frac{\lambda_1 + \lambda_2 + \dots + \lambda_n}{n} \quad (4-26.1)$$

характеризует положение центра тяжести всех, вообще говоря,

¹⁾ Этот вопрос возникает и во многих других физических задачах, а не только в теории электрических цепей.

комплексных корней знаменателя (т. е. характеристических частот цепи). Так как все λ_i должны иметь отрицательные вещественные части для того, чтобы цепь была устойчива, то коэффициент q_{n-1} должен быть *положительным*. Мы можем приписать ему любое значение, введя масштабное преобразование

$$z_1 = \alpha z, \quad (4-26.2)$$

где через z_1 обозначена первоначальная, еще ненормированная переменная преобразования Лапласа, тогда как z означает ту окончательную переменную, с которой нам нужно оперировать. Масштабный множитель α мы подберем так, чтобы новый коэффициент q_{n-1} был равен n . Это значит, что новая единица времени выбрана таким образом, что центр тяжести характеристических частот определен точкой -1 . Преобразование (2) переменной z сопровождается преобразованием масштаба времени¹⁾.

$$t_1 = \frac{t}{\alpha}. \quad (4-26.3)$$

Мы рассмотрим четыре различных решения задачи обращения преобразования Лапласа. Имея в виду преодоление вычислительных затруднений, мы можем выбрать то или другое из этих решений. Но и с математической точки зрения каждое из этих решений имеет свои достоинства.

27. Обращение с помощью полиномов Лежандра

В нашем первом решении мы используем равноотстоящие *вещественные* значения переменной z . Введем преобразование

$$e^{-t} = \xi. \quad (4-27.1)$$

Ценность этого преобразования состоит в том, что оно переводит бесконечный интервал $(0, \infty)$ переменной t в конечный интервал $(0, 1)$ переменной ξ . Теперь мы рассматриваем $f(t)$ как функцию новой переменной ξ , отмечая это записью $f(\xi)^*$. Тогда преобразование Лапласа принимает вид

$$\mathcal{L}(z) = \int_0^1 f(\xi) \xi^{z-1} d\xi. \quad (4-27.2)$$

¹⁾ Подстановка z из формулы (2) в $\mathcal{L}(z_1)$ уже предполагает, что преобразование Лапласа есть инвариант линейных преобразований. Фактически же преобразование дифференциала dt_1 требует, чтобы $\mathcal{L}(z)$ было разделено на α . Так как, однако, в задачах теории электрических цепей $f(t)$ означает единичный импульс, а этот импульс изменяется в силу преобразования масштаба на такой же множитель, то мы вправе принять, что преобразование Лапласа инвариантно относительно масштабного преобразования (2).

*) Сохранив обозначение f и для новой функции.

Рассмотрим ряд равноотстоящих точек

$$z = 1, 2, 3, \dots \quad (4-27.3)$$

и «моменты» функции $f(\xi)$:

$$y_k = \mathcal{L}(k+1) = \int_0^1 f(\xi) \xi^k d\xi. \quad (4-27.4)$$

Пусть $p_n(\xi)$ есть какой-нибудь полином относительно ξ :

$$p_n(\xi) = p_n^0 + p_n^1 \xi + \dots + p_n^n \xi^n. \quad (4-27.5)$$

Если каждому моменту y_k мы припишем вес, равный соответствующему коэффициенту этого полинома

$$c_n = \sum_{\alpha=0}^n y_\alpha p_n^\alpha, \quad (4-27.6)$$

то получим определенный интеграл

$$c_n = \int_0^1 f(\xi) p_n(\xi) d\xi. \quad (4-27.7)$$

Мы воспользуемся обозначением $p_n(y)$ в следующем смысле: выпишем полином $p_n(y)$ и заменим y^k через y_k . При таком соглашении формула (7) может быть записана в следующей условной форме:

$$c_n = p_n(y). \quad (4-27.8)$$

Теперь мы введем полиномы Лежандра $P_n(x)$ [ср. (5-20.11)], нормировав, однако, их интервал *определения* к интервалу $[0, 1]$, вместо обычного интервала $[-1, 1]$. Эти полиномы тогда непосредственно выражаются через гипергеометрическую функцию

$$P_k^*(x) = F(-k, k+1, 1; x); \quad (4-27.9)$$

их «норма» равна

$$N_k = \int_0^1 P_k^{*2}(x) dx = \frac{1}{2k+1}. \quad (4-27.10)$$

Мы можем разложить нашу функцию $f(\xi)$ по этим полиномам [ср. (5-16.9)]:

$$f(\xi) = \sum_{k=0}^{\infty} (2k+1) c_k P_k^*(\xi), \quad (4-27.11)$$

где

$$c_k = \int_0^1 f(\xi) P_k^*(\xi) d\xi = P_k^*(y). \quad (4-27.12)$$

Таким образом, мы получим явное представление функции $f(\xi)$ в форме сходящегося бесконечного ряда, коэффициенты которого могут быть выражены через y_k , т. е. через равноотстоящие значения преобразования Лапласа вдоль вещественной оси.

Коэффициенты полиномов $P_n^*(\xi)$ обладают тем ценным свойством, что все они суть *целые числа*; они могут быть заранее вычислены (ср. табл. IV) и представлены в форме треугольной матрицы. Умножение заданных y_k на эту матрицу дает коэффициенты c_k ; умножая затем эти коэффициенты на $(2k+1)$, получим коэффициенты разложения (12). К несчастью, элементы этой матрицы возрастают очень быстро. Поэтому для того, чтобы коэффициенты c_k были вычислены хотя бы с умеренной точностью, значения y_k должны быть заданы с *исключительной точностью*. При $k=10$ элементы матрицы возрастают до $2 \cdot 10^6$. Поэтому идти дальше $k=10$ не имеет смысла, даже если полностью использовать точность десяти значащих цифр обыкновенных арифмометров. Точность, с которой должны быть заданы y_k , составляет уже здесь десять значащих цифр. Измерения далеко не достигают такой точности, и это показывает, что физические *наблюдения* преобразования Лапласа никогда не могут привести к решению задачи восстановления исходной функции с достаточной степенью точности. Соотношение между заданными величинами и первоначальной функцией исключительно «косоугольно», и малейшее «искажение» данных разрушает всякую надежду на построение обращения эмпирически заданного преобразования Лапласа.

Здесь мы сталкиваемся с теми же затруднениями, которые встретились при разложении функции по показательным функциям (ср. § 23). Пример этот не случаен, так как функцию этого типа можно действительно рассматривать как преобразование Лапласа некоторой функции.

Однако положение будет совершенно иным, когда $\mathcal{L}(z)$ есть *теоретически* заданная функция, например отношение двух полиномов, характерное для задач теории электрических цепей:

$$\mathcal{L}(z) = \frac{p(z)}{q(z)}.$$

Для получения величин y_k нужно подставить в два полинома последовательные целые числа 1, 2, 3, ... (это делается быстро и просто). Коэффициенты этих полиномов не должны быть заданы больше, чем с минимальным числом значащих цифр (2 или 3). Но во всех последующих расчетах требуется точность всех десяти значащих цифр. Действительно, если нужно получить более одиннадцати членов разложения (при $n=10$), необходимо, чтобы основные данные были заданы с 15 десятичными знаками. Это не приводит к затруднениям, так как $p(k)$, $q(k)$ при подстановке последовательных целых чисел

редко будут содержать более десяти значащих цифр даже при соблюдении абсолютной точности. Следовательно, необходимо лишь выполнить деление с точностью до пятнадцати значащих цифр, для чего потребуется поставить снова на арифмометр остаток от первого деления и произвести деление этого остатка на прежний знаменатель. Кроме того, коэффициенты полиномов $P_n^*(x)$ возрастают не свыше, чем до 10 разрядов вплоть до $n=14$ и поэтому не требуют удвоенной точности. Если масштабный множитель выбран надлежащим образом, редко потребуется больше чем пятнадцать членов разложения.

28. Обращение с помощью полиномов Чебышева

Хотя полиномы Лежандра $P_k(x)$ табулированы, можно все же получить более удобное разложение в форме *ряда Фурье*. Введем угловую переменную θ , полагая

$$e^{-t} = \xi = \frac{1 + \cos \theta}{2} = \cos^2 \frac{\theta}{2}. \quad (4-28.1)$$

Будем считать, что $f(\xi)$ есть функция от угла θ и определена в интервале $(0, \pi)$; разложим ее в ряд Фурье по синусам

$$f(\theta) = \frac{4}{\pi} (b_1 \sin \theta + b_2 \sin 2\theta + \dots). \quad (4-28.2)$$

Проверим, выполняются ли условия на концах: $f(0) = f(\pi) = 0$. Точка $\theta = \pi$ соответствует $t = \infty$, а там импульсная реакция равна нулю. С другой стороны, в точке $\theta = 0$, т. е. в точке $t = 0$, функция, вообще говоря, не равна нулю. Мы можем, однако, с помощью простого приема сделать ее равной нулю. Если $\mathcal{L}(z)$ имеет в бесконечности вид $\frac{p_1}{z}$ [ср. (24.8)], то мы положим

$$f_1(t) = f(t) - p_1 e^{-t} \quad (4-28.3)$$

и

$$\mathcal{L}_1(z) = \mathcal{L}(z) - \frac{p_1}{z+1}. \quad (4-28.4)$$

Тогда $\mathcal{L}_1(z)$ убывает на бесконечности как z^{-2} , и функция $f_1(t)$ равна нулю при $t=0$. Естественные условия на концах для разложения (2) теперь удовлетворяются, и сходимость ряда будет удовлетворительной.

Коэффициенты b_k ряда (2) определяются обычным образом:

$$b_k = \frac{1}{2} \int_0^\pi f_1(\theta) \sin k\theta d\theta = - \int_0^1 f_1(\xi) \frac{\sin k\theta}{\sin \theta} d\xi. \quad (4-28.5)$$

Но функция $\frac{\sin k\theta}{\sin \theta}$, рассматриваемая как функция от ξ , есть полином Чебышева второго рода (ср. III, 4 и V, 20):

$$\frac{\sin k\theta}{\sin \theta} = U_{k-1}^*(\xi) = \frac{1}{2k} T_k^*(x). \quad (4-28.6)$$

Коэффициенты этого полинома снова суть целые числа, и опять можно установить прямую связь полинома с гипергеометрическим рядом:

$$U_{k-1}^*(\xi) = (-1)^{k-1} kF(k+1, -k+1, \frac{3}{2}; \xi). \quad (4-28.7)$$

Применив к нашему случаю формулы (27.7), (27.8), получим

$$b_k = U_{k-1}^*(y). \quad (4-28.8)$$

Коэффициенты этих полиномов могут быть заранее вычислены (см. табл. IX). Они также могут быть представлены треугольной матрицей. Умножение основных значений y_k на элементы этой матрицы опять порождает коэффициенты ортогонального разложения, аналогично предыдущему разложению, с тем дополнительным преимуществом, что получающийся ряд может быть записан как ряд Фурье относительно угловой переменной θ . Здесь снова следует соблюдать крайнюю точность при числовых выкладках, и мы не можем идти дальше $n=10$, если y_k не задаются с точностью, большей чем 10 разрядов. Коэффициенты полиномов $U_{k-1}^*(\xi)$ становятся десятизначными при $k=14$.

29. Обращение с помощью рядов Фурье

Весьма неортогональные свойства равноотстоящих точек функции $\mathcal{L}(z)$ на вещественной оси переходят в совершенно другие свойства, если рассматривать равноотстоящие точки на *мнимой оси*. Мы воспользуемся тем фактом, что в задачах теории связи искомая функция (импульсная реакция) практически существует лишь в конечном интервале, так как вне времени запоминания T_0 импульсная реакция становится пренебрежимо малой. Этот факт дает возможность указать такое решение задачи обращения, при котором используются лишь равноотстоящие значения функции $\mathcal{L}(z)$ вдоль мнимой оси.

Мы начнем свое исследование следующим воображаемым экспериментом. Вместо того, чтобы осуществить единичный импульс только один раз, мы станем повторять его ритмически через интервал T_0 или больший, чем T_0 . Тогда импульсная реакция также станет повторяться ритмически, причем последовательные реакции не будут накладываться друг на друга, так как промежуток времени, их разделяющий, достаточен для того, чтобы реакция от предыдущего периода заглохла.

Мы имеем теперь ситуацию, при которой как $f(t)$, так и $g(t)$ суть периодические функции с периодом

$$T \geq T_0.$$

Мы положим

$$\omega_0 = \frac{2\pi}{T_0}. \quad (4-29.1)$$

Входную функцию, если разложить ее на гармонические составляющие, можно записать следующим образом (ряд этот может и расходиться):

$$\begin{aligned} f(t) &= \frac{2}{T} \left[\frac{1}{2} + \cos \frac{2\pi}{T} t + \cos 2 \frac{2\pi}{T} t + \dots \right] = \\ &= \frac{\omega_0}{\pi} \left[\frac{1}{2} + \cos \omega_0 t + \cos 2\omega_0 t + \dots \right]. \end{aligned} \quad (4-29.2)$$

Соответствующая функция выхода $g(t)$ принимает вид (если мы вспомним определение (18.4) для частотной реакции: $\psi(\omega) = \mathcal{L}(i\omega)$):

$$g(t) = K(t) = \frac{\omega_0}{\pi} \sum_{k=0}^{\infty} [\mathcal{L}(ik\omega_0) e^{ik\omega_0 t} + \mathcal{L}(-ik\omega_0) e^{-ik\omega_0 t}], \quad (4-29.3)$$

или, если разделить вещественную и мнимую части функции,

$$\mathcal{L}(i\omega) = A(\omega) + iB(\omega),$$

$$K(t) = \frac{\omega_0}{\pi} \left[\frac{A(0)}{2} + \sum_{k=1}^{\infty} A(k\omega_0) \cos k\omega_0 t - \sum_{k=1}^{\infty} B(k\omega_0) \sin k\omega_0 t \right]. \quad (4-29.4)$$

Мы имеем здесь разложение $K(t)$ в ряд Фурье, использующее равноотстоящие значения преобразования Лапласа вдоль мнимой оси. Если введем новую переменную z_1 , положив $z = i\omega z_1$, и внесем ее в полиномы $p(z)$ и $q(z)$, то снова придется выполнить только простую работу — подставить целые числа $0, 1, 2, \dots$ в два полинома и вычислить отношения этих полиномов. Единственное неудобство состоит в том, что коэффициенты полиномов будут теперь попеременно вещественными и мнимыми. Результатом подстановки будет комплексное число, и вычисление закончится делением одного комплексного числа на другое. Вещественные части этих частных дают коэффициенты ряда косинусов, а мнимые части — коэффициенты ряда синусов, взятые со знаком минус.

Два затруднения все же остаются. Сходимость полученного ряда может оказаться слишком слабой. Кроме того, мы можем заранее не знать значения времени запоминания T_0 . Однако эти затруднения можно преодолеть. Сходимость ряда можно ускорить, видоизменив

надлежащим образом функцию $K(t)$. С этой целью мы разлагаем сначала $\mathcal{L}(z)$ в окрестности точки $z = \infty$. Мы знаем заранее, что для больших значений z

$$\mathcal{L}(z) = \frac{p_1}{z} + \frac{p_2}{z^2} + \frac{p_3}{z^3}. \quad (4-29.5)$$

Это значит, что члены разложения Фурье (4) убывают слишком медленно, а именно как z^{-1} . Исправим теперь функцию $\mathcal{L}(z)$, положив

$$\mathcal{L}_1(z) = \mathcal{L}(z) - \frac{p_1}{z+1} - \frac{p_2 + p_1}{(z+1)^2}; \quad (4-29.6)$$

тогда для больших по модулю значений переменной z уменьшение происходит по закону z^{-3} , и сходимость ряда для $K(t)$ становится удовлетворительной. Мы можем остановиться в ряду на надлежаще выбранном k , если проследим на мнимой оси максимум функции $|\mathcal{L}_1(i\omega)|$ и затем отбросим члены, начиная с той точки ω_l , на которой соответствующее $|\mathcal{L}_1(i\omega_l)|$ снизилось до нескольких процентов от максимума.

Теперь надо решить вопрос о выборе частоты ω_0 , определенной формулой (1). Важно лишь, чтобы мы не перешли предела $\frac{2\pi}{T_0}$, тогда как меньшие значения нам не мешают. Если мы примем, что

$$\omega_0 = \frac{1}{12} \omega_l, \quad (4-29.7)$$

то получим достаточно надежную гарантию того, что ω_0 не переоценено. Деление ω_l на 12 частей означает, что для представления $K(t)$ с точностью до нескольких процентов достаточно двенадцати членов с синусами и тринадцати членов с косинусами. Мы редко впадем в ошибку при таком предположении, так как ряд Фурье с 25 членами достаточно гибок для того, чтобы представить даже весьма негладкие функции. Построенный ряд представляет не первоначальную функцию $K(t)$, но следующую ее модификацию:

$$K_1(t) = K(t) - p_1 e^{-t} - (p_2 + p_1) t e^{-t}. \quad (4-29.8)$$

Эта функция начинается с нуля и имеет в нулевой точке горизонтальную касательную:

$$K_1(0) = 0, \quad K'_1(0) = 0. \quad (4-29.9)$$

Эти условия на границе и создают хорошую сходимость ряда Фурье. Если набросать ход кривой $K_1(t)$, вычисленной на основе ряда (9), то мы можем проверить годность нашего выбора (7), обнаружив, что функция $K_1(t)$ выпрямляется и становится практически равной нулю еще до того, как t достигает верхнего предела $\frac{2\pi}{\omega_0}$.

30. Обращение с помощью функций Лагерра

Первые два решения задачи обращения применимы к любому преобразованию Лапласа. Они используют равноотстоящие значения функции $\mathcal{L}(z)$ вдоль вещественной оси. Третье решение было специально приспособлено к требованиям анализа электрических цепей. Мы воспользовались тем фактом, что интервал интегрирования, хотя теоретически и простирающийся до бесконечности, фактически может быть сведен к конечному промежутку времени T_0 , — «времени запоминания» цепи. В настоящем параграфе рассматривается метод, который снова применим в общем случае. Он основан на теореме взаимности Фурье, которая решает задачу обращения преобразования Фурье. Обращение преобразования Лапласа может быть приведено к этой задаче.

В § 24 мы уже кратко коснулись соотношения между преобразованиями Лапласа и Фурье. Если рассматривать преобразование Лапласа $\mathcal{L}(z)$ для чисто мнимых значений $i\omega$ переменной z , то получается преобразование Фурье исходной функции $f(t)$. Мы можем теперь воспользоваться теоремой взаимности Фурье (17.16) и получить $f(t)$ как преобразование Фурье для функции $\mathcal{L}(i\omega)$:

$$\mathcal{L}(i\omega) = \int_0^{\infty} f(t) e^{-i\omega t} dt, \quad (4-30.1)$$

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \mathcal{L}(i\omega) e^{i\omega t} d\omega = \frac{1}{2\pi i} \int_{-i\infty}^{+i\infty} \mathcal{L}(z) e^{zt} dz. \quad (4-30.2)$$

Хотя мы, таким образом, получили $f(t)$ в виде определенного интеграла, действительное вычисление его отнюдь не тривиально ввиду бесконечных пределов интегрирования. Мы укажем сейчас метод, который даст возможность численно получить преобразование Фурье для функции, заданной либо аналитически, либо численно. С этой целью мы воспользуемся замечательным конформным отображением комплексной плоскости на себя, которое встретилось нам ранее в главе об алгебраических уравнениях (ср. I, 18). Это преобразование отображает всю правую комплексную полуплоскость на внутренность единичной окружности. Сама единичная окружность при этом оказывается образом бесконечной мнимой оси. Следовательно, интегрирование вдоль мнимой оси переходит в интегрирование вдоль единичной окружности.

Заменим комплексную переменную z в преобразовании Лапласа новой переменной v согласно следующим формулам:

$$z = \frac{1-v}{1+v} = -1 + \frac{2}{1+v},$$

$$dz = -\frac{2dv}{(1+v)^2} = -\frac{(z+1)^2}{2} dv. \quad (4-30.3)$$

В новой переменной интеграл (2) берется по единичной окружности

$$v = e^{i\theta}. \quad (4-30.4)$$

Интервал изменения угловой переменной θ есть $(-\pi, \pi)$. Поэтому

$$f(t) = \frac{1}{2\pi i} \int_{-i\infty}^{+i\infty} \mathcal{L}(z) e^{tz} dz = -\frac{1}{4\pi} \int_{-\pi}^{+\pi} \mathcal{L}(z) (1+z)^2 e^{tz+i\theta} d\theta. \quad (4-30.5)$$

Введем теперь функцию

$$\mathcal{L}_1(z) = \frac{(1+z)^2}{z} \left[\mathcal{L}(z) - \frac{p_{n-1}}{z+1} \right]. \quad (4-30.6)$$

Мы видели [ср. (25.2)], что степень полинома $p(z)$ ниже степени $q(z)$, вообще говоря, на единицу. Видоизменение функции $\mathcal{L}(z)$ путем вычитания последнего члена формулы (6) имеет своей целью повысить разность между степенями числителя и знаменателя — после приведения обеих дробей к одному знаменателю — до двух. Тогда умножение на $(1+z)^2$ уравнивает степени числителя и знаменателя, и функция $\mathcal{L}_1(z)$ остается конечной даже при $z = \infty$. В согласии с видоизменением функции $\mathcal{L}(z)$ мы положим

$$f(t) = p_{n-1} e^{-t} + f_1(t). \quad (4-30.7)$$

Функция $f_1(t)$ теперь имеет начальное значение нуль при $t = 0$.

Теперь мы преобразуем $\mathcal{L}_1(z)$ к новой переменной v . Это можно выполнить умножением коэффициентов числителя и знаменателя на табличные матрицы B_n (ср. табл. II). Такое преобразование выполняется легко, ибо элементы этих матриц все целые числа. Запись $\mathcal{L}_1(v)$ обозначает, что эта операция уже произведена над функцией $\mathcal{L}_1(z)$. (Если случится, что числитель в $\mathcal{L}_1(z)$ имеет меньшую степень, чем знаменатель, то мы все же рассматриваем числитель как полином n -й степени, заменяя недостающие коэффициенты нулями.)

Так как $\mathcal{L}_1(e^{i\theta})$ есть периодическая функция от θ , мы, естественно, разложим ее в ряд Фурье:

$$\mathcal{L}_1(e^{i\theta}) = \sum_{k=-\infty}^{\infty} (c_k e^{ik\theta} + c_{-k} e^{-ik\theta}) = \sum_{k=-\infty}^{\infty} (c_k v^k + c_{-k} v^{-k}). \quad (4-30.8)$$

Если речь идет о преобразовании Фурье для эмпирически заданной функции, то коэффициенты c_k , c_{-k} этого разложения могут быть получены с помощью тригонометрического интерполирования (ср. §§ 12, 13). Тогда представление тригонометрическим рядом будет, конечно, неточным. Однако достаточно знать максимум абсолютной ошибки ε этого разложения в любой точке единичной окружности.

Мы можем теперь оценить максимальную ошибку преобразования Фурье (5), если $f(t)$ видоизменена так, как это указано в формуле (7). Интеграл (5) показывает, что $\mathcal{L}_1(z)$ умножается вдоль пути интегрирования на множитель, абсолютная величина которого равна 1, независимо от того, какое значение имеет t (при t вещественном). Следовательно, мы можем считать, что максимальная абсолютная погрешность функции $f_1(t)$, вычисленной на основе формулы (5), не может превысить

$$|r_1(t)| \leq \varepsilon. \quad (4-30.9)$$

В случае обращения преобразования Лапласа можно утверждать гораздо большее. Здесь мы знаем, что $\mathcal{L}_1(v)$ есть аналитическая функция от v , не имеющая особых точек ни внутри, ни на единичной окружности. Мы можем разложить $\mathcal{L}_1(v)$ в ряд Тейлора в окрестности точки $v=0$:

$$\mathcal{L}_1(v) = c_0 + c_1 v + c_2 v^2 + \dots \quad (4-30.10)$$

Это разложение сходится всюду внутри и на единичной окружности. Следовательно, здесь мы находимся в том благоприятном положении, что можем найти точные значения коэффициентов Фурье для $\mathcal{L}_1(v)$ вдоль единичной окружности без какого-либо интегрирования. Сравнение с общей формой разложения (8) показывает, что коэффициенты c_{-k} совсем отсутствуют. Это соответствует тому факту, что $f(t)$ тождественно равно нулю для всех значений t , меньших нуля.

В задачах теории электрических цепей коэффициенты можно получить с помощью простого и изящного алгоритма, «алгебраического деления двух полиномов», который мы встречали раньше при изучении корней алгебраического уравнения (ср. I, 14). Единственное отличие состоит в том, что в нашей прежней схеме разложение производилось по убывающим степеням переменной, тогда как теперь требуется разложение по возрастающим степеням переменной. Но это только значит, что коэффициенты полиномов $p(v)$ и $q(v)$ записаны в обратном порядке, т. е. начиная со свободного члена и кончая высшей степенью. Свободный член q_0 знаменателя не может быть нулем, ибо $q(v)$ не может иметь корня внутри единичного круга и уж заведомо не имеет корня в его центре $v=0$. Поэтому всегда можно разделить числитель и знаменатель на q_0 , сделав свободный член знаменателя равным единице. Записываем последовательные коэффициенты числителя (расположенного по возрастающим степеням) на «неподвижную полосу», а последовательные коэффициенты знаменателя,—поднимаясь снизу вверх и меняя знаки всех коэффициентов, кроме первого,—на «подвижную полосу». Тогда обычным приемом подвижной полосы (I, 8) получим на нарастающей полосе последовательные коэффициенты. Этот алгоритм продолжим до получения пятнадцати — двадцати коэффициентов.

Наша задача теперь свелась к обращению специального преобразования Лапласа:

$$\frac{2v^k}{(1+z)^2} = \frac{2(1-z)^k}{(1+z)^{k+2}} = \frac{2(-1)^k z_1^k}{(2+z_1)^{k+2}}, \quad (4-30.11)$$

где мы положили

$$z = 1 + z_1. \quad (4-30.12)$$

Мы решим эту задачу в несколько этапов. Прежде всего, введя z_1 в качестве новой переменной, получим

$$\mathcal{L}(z_1) = \int_0^\infty f(t) e^{-t} e^{-z_1 t} dt = \int_0^\infty \varphi(t) e^{-z_1 t} dt, \quad (4-30.13)$$

где

$$\varphi(t) = f(t) e^{-t}. \quad (4-30.14)$$

Но мы знаем, что исходная функция для преобразования $(2+z)^{-1}$ есть

$$f(t) = e^{-2t}.$$

Известно также, что умножение исходной функции на t соответствует дифференцированию преобразования с изменением знака. Поэтому преобразованием исходной функции

$$t^{k+1} e^{-2t}$$

будет

$$\mathcal{L}(z_1) = \frac{(k+1)!}{(2+z_1)^{k+2}}.$$

Аналогично умножение преобразования на z_1 соответствует дифференцированию исходной функции. Поэтому исходной функцией для преобразования

$$\mathcal{L}(z_1) = z_1^{k+1} \frac{(k+1)!}{(2+z_1)^{k+2}} \quad (4-30.15)$$

будет

$$\frac{d^{k+1}}{dt^{k+1}} (t^{k+1} e^{-2t}). \quad (4-30.16)$$

Важный класс ортогональных полиномов, называемых «полиномами Лагерра», определяется следующей формулой *):

$$\mathcal{L}_n(t) = e^t \frac{d^n}{dt^n} (t^n e^{-t}). \quad (4-30.17)$$

*) См. {1}, стр. 93.

Отсюда следует, что

$$\mathcal{L}_n(2t) = e^{2t} \frac{d^n}{dt^n} (t^n e^{-2t}). \quad (4-30.18)$$

На основании формул (15) — (18) можно утверждать, что исходная функция —

$$\varphi(t) = \frac{e^{-2t}}{(k+1)!} \mathcal{L}_{k+1}(2t),$$

а преобразование Лапласа имеет вид

$$\mathcal{L}(z_1) = \frac{z_1^{k+1}}{(2+z_1)^{k+2}}.$$

Далее, исходная функция —

$$\varphi(t) = e^{-2t} \left[\frac{L_k(2t)}{k!} - \frac{L_{k+1}(t)}{(k+1)!} \right],$$

преобразование Лапласа —

$$\mathcal{L}(z_1) = \frac{z_1^k}{(2+z_1)^{k+1}} - \frac{z_1^{k+1}}{(2+z_1)^{k+2}} = \frac{2z_1^k}{(2+z_1)^{k+2}}.$$

Возвращаясь к первоначальной переменной z , мы должны умножить $\varphi(t)$ на e^t . Таким образом, мы приходим к заключению, что обращением преобразования Лапласа (11) является

$$(-1)^k e^{-t} \left[\frac{L_k(2t)}{k!} - \frac{L_{k+1}(2t)}{(k+1)!} \right]. \quad (4-30.19)$$

Введем теперь так называемые «функции Лагерра» [ср. (5-20.6)]:

$$\varphi_k(t) = \frac{e^{-\frac{t}{2}} L_k(t)}{k!}, \quad (4-30.20)$$

которые образуют систему ортонормальных функций в интервале $(0, \infty)$ переменной t . Выражение (19) может быть тогда записано в форме

$$(-1)^k [\varphi_k(2t) - \varphi_{k+1}(2t)],$$

а обращением ряда (10) будет

$$\begin{aligned} f_1(t) &= c_0 \varphi_0(t) - (c_0 + c_1) \varphi_1(t) + (c_1 + c_2) \varphi_2(t) - \dots = \\ &= \sum_{k=0}^{\infty} (-1)^k (c_{k-1} + c_k) \varphi_k(2t). \end{aligned} \quad (4-30.21)$$

Как мы уже видели раньше, замена этого ряда конечным разложением приводит к погрешности, которая ни при каком значении t не может превысить погрешность усеченного ряда (10).

Для функций Лагерра $\psi(2t)$ можно заранее составить таблицы через целесообразно выбранные малые интервалы переменной t и до нужного значения k (например, до $k=20$)¹⁾.

Затем, подставив вместо аргумента t достаточно плотную последовательность его значений, мы получим результирующее разложение (21). Надо при этом не забыть добавить поправку (7), чтобы восстановить полную импульсную реакцию сети. Эта поправка сводится просто к тому, что коэффициент функции $\varphi_0(2t) = e^{-t}$ заменяется следующим выражением:

$$c'_0 = c_0 + p_{n-1}. \quad (4-30.22)$$

Подводя итог, мы можем сказать, что получено четыре удобных численных схемы для построения импульсной реакции сети, заданной своими структурными элементами, без вычисления корней полинома $q(z)$. В то время как первые два приема используют в качестве узловых равноотстоящие значения функции $\mathcal{L}(z)$ вдоль вещественной оси, третий прием основывается на равноотстоящих значениях вдоль мнимой оси, а метод настоящего параграфа использует бесконечно малую окрестность одной точки $z=1$. Значение функции $\mathcal{L}(z)$ и всех ее производных при $z=1$ однозначно определяет коэффициенты c_k разложения (11), а следовательно, и исходной функции $f(t)$.

Во всех этих схемах должно выполняться одно фундаментальное условие. Электрическая цепь не должна содержать высокочастотных колебаний со слабым затуханием. Надлежащее представление таких колебаний в наших разложениях потребовало бы непомерно большого числа членов. Однако колебание этого рода соответствует корню, очень близкому к мнимой оси. Такие корни без особого труда можно выявить (ср. I, 19) и отдельно определить их влияние. Надо только перед проведением нашей численной схемы обращения преобразования Лапласа вычесть ту часть, которая соответствует этим корням [ср. (25.5)].

31. Интерполяция преобразования Лапласа

Преобразование Лапласа $\mathcal{L}(z)$ как аналитическая функция комплексного переменного z однозначно определено, если оно задано на произвольно малой непрерывной дуге комплексной плоскости. Однако более интересен метод задания аналитической функции на бесконечно большом числе равноотстоящих точек и определения всех других ее значений путем интерполяции и экстраполяции. Такую интерполяцион-

¹⁾ Такую таблицу нормированных функций Лагерра составила Фанни Гордон, сотрудница Национального Бюро Стандартов, в период работы автора в Институте численного анализа Национального Бюро Стандартов в Лос-Анжелосе, Калифорния. Эта таблица приведена в приложении (см. табл. X).

ную формулу для преобразования Лапласа, действительно, можно составить.

Мы видели в § 27, что, задавая преобразование Лапласа в целых точках $1, 2, 3, \dots$ вдоль вещественной оси, мы можем однозначно определить «исходную функцию» $f(\xi)$, которая преобразована в $\mathcal{L}(z)$ по формуле (27.4). Подставим теперь в (27.4) бесконечный ряд (27.11) для $f(\xi)$ и проинтегрируем почленно. При этом нам понадобятся определенные интегралы

$$w_k(z+1) = \int_0^1 \xi^z P_k^*(\xi) d\xi = \sum_{\alpha=0}^k \frac{p_k^\alpha}{z+\alpha+1}. \quad (4-31.1)$$

Если привести эти дроби к одному знаменателю, то получим в числителе полином степени k , а в знаменателе произведение

$$(z+1)(z+2)(z+3)\dots(z+k+1).$$

Числитель также может быть разложен на произведение корневых двучленов, и из того факта, что каждая целая степень $z=0, 1, 2, \dots, k-1$ ортогональна к $P_k^*(\xi)$, можно заключить, что это произведение должно равняться

$$z(z-1)(z-2)\dots(z-k+1).$$

Остается определить лишь постоянный множитель, предшествующий этому произведению. Заставляя z стремиться к бесконечности и учитывая, что $P_k^*(1)=1$, находим, что искомая постоянная равна 1. Вводя символ

$$\left[\begin{matrix} z \\ k \end{matrix} \right] = \frac{z(z-1)(z-2)\dots(z-k+1)}{(z+1)(z+2)\dots(z+k+1)} = \frac{(z!)^2}{(z+k+1)!(z-k)!}, \quad (4-31.2)$$

мы получаем

$$w_k(z+1) = \left[\begin{matrix} z \\ k \end{matrix} \right], \quad (4-31.3)$$

и интерполяционная формула для преобразования Лапласа приводится к виду

$$\mathcal{L}(z+1) = \sum_{k=0}^{\infty} c_k \left[\begin{matrix} z \\ k \end{matrix} \right] = \sum_{k=0}^{\infty} (2k+1) P_k^*(y) \left[\begin{matrix} z \\ k \end{matrix} \right]. \quad (4-31.4)$$

Символическое обозначение $P_k^*(y)$ (ср. § 27) имеет здесь следующий смысл: мы образуем полином $P_k^*(y)$ и заменяем y^α величиной $\mathcal{L}(\alpha+1)$, т. е. значением функции $\mathcal{L}(z)$ в целочисленной точке $z=\alpha+1$. Коэффициенты интерполяционной формулы (4), таким образом, опре-

деляются значениями функции $\mathcal{L}(z)$ в точках $z = 1, 2, 3, \dots$. Если выполняется условие

$$\sum_{k=0}^{\infty} |c_k| = \text{конечному числу}, \quad (4-31.5)$$

то ряд (4) будет сходиться для всех точек z , расположенных вправо от мнимой оси.

Хотя эта интерполяционная формула устанавливает тот факт, что преобразование Лапласа однозначно определяется, если известны его значения для всех целых точек, было бы ошибкой считать, что эти значения могут быть выбраны по произволу. Прежде всего из произвольности масштаба мы заключаем, что «единичное расстояние» может быть растянуто до любой длины: следовательно, мы могли бы разредить наши первоначальные данные до какой угодно степени.

Кроме того, преобразование

$$z = \alpha + z_1, \quad f(x) = f_1(x) e^{\alpha x} \quad (4-31.6)$$

показывает, что точки $z_1 = 1, 2, 3, \dots$ соответствуют точкам $z = \alpha + 1, \alpha + 2, \alpha + 3, \dots$. Следовательно, мы не только можем сколь угодно раздвинуть наши данные, но и начать их с точки, произвольно далекой от начала. Наши функциональные данные являются, таким образом, всегда в бесконечной степени избыточными, и мы никогда не можем свести их к основной системе, которая была бы необходима и достаточна. Это как будто противоречит тому факту, что коэффициенты c_k всегда могут быть образованы для произвольно заданного множества значений $\mathcal{L}(m)$ ($m = 1, 2, 3, \dots$). Однако, если эти значения заданы не надлежащим образом, то сумма (5) не будет сходиться, и наша интерполяционная формула потеряет смысл.

Из интерполяционной формулы (4) можно вывести и другое заключение. Преобразование, соответствующее [произвольной] четырех-полюсной цепи, всегда есть отношение двух полиномов. Такого рода преобразование является лишь весьма специальным классом преобразований Лапласа, характеризующихся такими исходными функциями, которые определяются как конечные линейные комбинации показательных функций с показателями, имеющими отрицательную вещественную часть. Однако сходимость разложения (4) означает, что произвольное преобразование Лапласа может быть порождено с любой степенью точности с помощью электрических цепей, а в действительности даже с помощью чисто «баллистических» цепей без всякой самоиндукции L . В самом деле, любая абсолютно интегрируемая функция с ограниченной вариацией может быть аппроксимирована конечным разложением Лежандра вида (27.11) с погрешностью, которую можно сделать сколь угодно малой. Соответствующее преобразование Лапласа (4) может быть записано как отношение двух полиномов,

а корни знаменателя можно привести к $z = -1, -2, -3, \dots$. Таким образом, мы не только можем имитировать произвольно заданную импульсную реакцию $K(t)$ с помощью RC -цепей, но можем нормировать константы этих цепей таким образом, чтобы произведения RC находились в отношении $1 : \frac{1}{2} : \frac{1}{3} : \frac{1}{4} : \dots$ друг к другу. Особенно поучительно приложение интерполяционной формулы (4) — если ограничиться n членами разложения — к функции

$$\frac{1}{(z + \beta)^2 + \gamma^2},$$

явно показывающее, что выход одной RCL -цепи может быть заменен выходом системы достаточного числа надлежаще соединенных цепей, хотя выход каждой отдельной из этих цепей есть монотонно убывающая функция без каких-либо колебаний.

Переходим теперь ко второму методу обращения, изученному в § 28. Здесь исходная функция $f(\xi)$ была разложена в ряд Фурье по синусам; коэффициенты этого ряда были получены с помощью той же функции $\mathcal{L}(m+1)$, как и раньше, но путем умножения на другой ряд весовых множителей (коэффициенты полиномов Чебышева вместо коэффициентов полиномов Лежандра). Теперь вместо интегралов (1) появляются интегралы

$$w_k(z+1) = \int_0^1 \xi^z \sin k\theta d\xi = \int_0^\pi \cos^{2z} \frac{\theta}{2} \sin k\theta \sin \theta d\theta. \quad (4-31.7)$$

Такой интеграл может быть выражен через гамма-функции, и мы получаем

$$w_k(z+1) = k \sqrt{\pi} \frac{\left(z + \frac{1}{2}\right)! z!}{(z+k+1)!(z-k+1)!}. \quad (4-31.8)$$

Мы снова вводим упрощающее обозначение

$$\left\{ \begin{matrix} z \\ k \end{matrix} \right\} = \frac{\left(z + \frac{1}{2}\right)! z!}{(z+k+1)!(z-k+1)!} \quad (4-31.9)$$

и получаем интерполяционную формулу

$$\mathcal{L}(z+1) = \frac{4}{\sqrt{\pi}} \sum_{k=1}^{\infty} k b_k \left\{ \begin{matrix} z \\ k \end{matrix} \right\} = \frac{4}{\sqrt{\pi}} \sum_{k=1}^{\infty} k U_{k-1}^*(y) \left\{ \begin{matrix} z \\ k \end{matrix} \right\}. \quad (4-31.10)$$

Полюсами функций (9) являются точки $z = -\frac{3}{2}, -\frac{5}{2}, -\frac{7}{2}, \dots$, но эти функции не могут быть представлены в виде отношения двух

полиномов и поэтому не могут быть реализованы с помощью электрических цепей.

Еще одну интерполяционную формулу можно вывести, если воспользоваться третьим методом обращения преобразования Лапласа (см. § 29). Этот метод применим лишь в случае конечного промежутка времени интегрирования T_0 , как это и имеет место в задачах теории электрических цепей. Здесь точками интерполяции служат равноотстоящие точки вдоль мнимой оси. Метод вывода интерполяционной формулы тот же, что и раньше: мы подставляем ряд, полученный для $f(t)$, в формулу, определяющую преобразование Лапласа, и интегрируем почленно. Таким образом получаем

$$\mathcal{L}(z) = 2e^{\frac{T_0 z}{2}} \sum_{k=-\infty}^{+\infty} \mathcal{L}(ik\omega_0) \frac{\sin h \frac{z T_0}{2}}{z T_0 - 2\pi i k}. \quad (4-31.11)$$

Базисные интерполирующие функции теперь не имеют никаких полюсов и не могут быть реализованы с помощью электрических цепей.

Наконец, в четвертом из указанных выше решений задачи обращения мы разлагали функцию $\mathcal{L}(z)$ в ряд в окрестности точки $z=1$ после замены z новой переменной v .

Полученное разложение может быть, однако, записано и относительно исходной переменной z , и мы получаем следующую формулу:

$$\mathcal{L}(z) = 2 \sum_{k=0}^{\infty} \frac{L_k(-2y)}{k!} \left(\frac{1-z}{1+z} \right)^k \frac{1}{1+z}, \quad (4-31.12)$$

где обозначение $L_k(-2y)$ употреблено в том смысле, что в полиноме Лагерра $L_k(-2y)$ следует заменить y^x через $\mathcal{L}^x(1)$. Этот ряд сходится для всех комплексных значений z , вещественная часть которых положительна.

Интерпретируя эту интерполяционную форму в терминах электрической цепи, мы видим, что произвольное преобразование Лапласа образовано с помощью баллистических цепей, произведения RC которых расположены в окрестности константы 1. Интерполирующие функции имеют полюс в точке -1 , но этот полюс имеет все возрастающий порядок k . А полюс кратности k в точке $z=-1$ эквивалентен суперпозиции k RC -цепей надлежаще выбранной интенсивности и с произведениями RC , отличающимися от 1 на произвольно малые величины $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_k$.

В этой интерполяционной (или, вернее, экстраполяционной) формуле узловые значения равны значениям функции $\mathcal{L}(z)$, взятым в бесконечно малой окрестности точки $z=1$. Можно получить еще другую интерполяционную формулу, взяв в качестве узловых значений все множество значений функции $\mathcal{L}(i\omega)$ вдоль

мнимой оси. С этой целью мы рассмотрим всю правую полуплоскость комплексной переменной z , отображенную внутрь единичной окружности переменной v . В этом новом отображении мы можем воспользоваться формулой Коши

$$\mathcal{L}(v) = \frac{1}{2\pi i} \oint \frac{\mathcal{L}(v_0) dv_0}{v - v_0}, \quad (4-31.13)$$

где v_0 пробегает единичную окружность. Возвращаясь к первоначальной переменной z , получаем

$$\mathcal{L}(z) = \frac{1}{2\pi} (1 + z) \int_{-\infty}^{+\infty} \frac{\mathcal{L}(i\omega) d\omega}{(1 + i\omega)(z - i\omega)}. \quad (4-31.14)$$

Так как интеграл есть предел суммы, то мы заключаем, что произвольная исходная функция — и порожденное ею преобразование Лапласа — может также быть имитирована только LC -цепями, не содержащими R , т. е. с помощью чисто синусоидальных колебаний без затухания. Мы имеем здесь крайнюю противоположность баллистическим цепям интерполяционной формулы (4): как будто бы узловые значения, расположенные вдоль вещественной оси, повернулись вокруг начала на 90° . Между этими двумя случаями существует фундаментальное отличие. Интенсивность имитирующих цепей в мнимом случае непрерывна и не имеет каких-либо особенностей в силу ортогональности базисных функций. В вещественном случае, однако, интенсивность имитирующих баллистических цепей должна быть выбрана особым образом в силу весьма неортогонального характера базисных интерполирующих функций.

ЛИТЕРАТУРА К ГЛАВЕ IV

- [1] Bush V., Operational Circuit Analysis (Wiley, New York, 1929).
- [2] Carslaw H. S., Theory of Fourier Series and Integrals (Macmillan, London, 1921).
- [2a] Толстов Г. П., Ряды Фурье, Физматгиз, М., 1960.
- [3] Churchill R. V., Fourier Series and Boundary Value Problems (McGraw-Hill, New York, 1941).
- [4] Franklin Ph., Fourier Methods (McGraw-Hill, New York, 1949).
- [5] Guillemin E. A., Mathematics of Circuit Analysis (Wiley, New York, 1949).
- [6] Jackson D., Fourier Series and Orthogonal Polynomials (Math. Assoc. of America, Oberlin, 1941).
- [6a] Геронимус Я. Л., Теория ортогональных многочленов, Гостехиздат, 1950.
- [7] Thomson W. T., Laplace Transformation (Prentice-Hall, New York, 1950).
- [7a] Лурье А. И., Операционное исчисление, М., 1950.
- [7б] Дёч Г., Руководство к практическому применению преобразования Лапласа, Физматгиз, М., 1960.
- [8] Widder D. V., The Laplace Transform (Princeton University Press, Princeton, 1941).

- [9] Wiener N., The Fourier Integral and Its Applications. (Cambridge University Press, New York, 1933).
[9a] Серебренников М. Г., Гармонический анализ, Гостехиздат, М., 1948.
[9б] Лопшиц А. М., Шаблоны для гармонического анализа, Гостехиздат, М., 1948.

Статьи:

- [10] Danielson G. C. and Lanczos C., Improvements in Practical Fourier Analysis and Their Application to X-ray Scattering from Liquids, J. Franklin Inst. 233, 365, 435 (1942).
[11] Bellman R., On the separation of exponentials, Boll. Unione Matem. Ital., 1960, III, 15, № 1, 38—39.
-

ГЛАВА V

АНАЛИЗ ЭМПИРИЧЕСКИХ ДАННЫХ

1. Историческое введение

Ограниченность точности физических наблюдений была осознана еще в древние времена. Архимед оценивал минимальные размеры вселенной, исходя из отсутствия сколько-нибудь заметного годичного параллакса в пределах погрешности тогдашних инструментов, считая истинной гелиоцентрическую теорию Аристарха. По мере того, как возрастала точность измерительных приборов, вырисовывалась определенная математическая проблема. Допустим, что мы имеем возможность наблюдать одно и то же явление многократно, при практически тождественных условиях. «Практически тождественные» означает, что решающие факторы, которые вызывают определенное физическое событие, остаются неизменными, в то время как не поддающиеся контролю второстепенные условия изменяются случайным образом. Например, мы можем многократно измерить зависимость между перемещением и промежутками времени движения шара, свободно падающего на землю. Масса шара и сила тяжести, определяющие это событие, остаются неизменными, тогда как небольшие несовершенства приборов измерения длины и времени беспорядочно меняются. Что, собственно, означают эти «беспорядочные» или «случайные» изменения в каждом отдельном случае, часто бывает трудно установить. Из рассмотрения такого рода задач возникает потребность в общем математическом приеме, при помощи которого можно было бы обработать избыточные измерения, имея в виду наилучшее использование всех измерений.

В начале XIX столетия Гаусс (1809) и Лежандр (1806) нашли независимо друг от друга замечательный универсальный метод, с помощью которого можно целесообразно использовать избыток данных измерения. Этот метод, известный как «метод наименьших квадратов», дал ответ на проблему избыточных данных, основанный в большей степени на чисто математической логике, чем на индуктивном физическом рассуждении. Позднее этот метод был развит и стал краеугольным камнем всех статистических исследований.

Задача анализа эмпирических данных имеет еще и другой важный аспект. Измерения производятся всегда в *дискретном* множестве

точек, между тем как функции, которые мы считаем лежащими в основе наших измерений, определены для *непрерывного* интервала переменной. По существу мы всегда *табулируем* функции — является ли это табулирование результатом простых вычислений или физических наблюдений. Табулирование приводит к задаче определения значений неизвестной функции в точках, расположенных *между* точками, приведенными в таблице. Это — та проблема «интерполяции», которая привлекала математическую мысль с давних времен. Великие аналитики XVII столетия были хорошо знакомы с методами равноотстоящей интерполяции полиномами и развили теорию конечных разностей до высокого уровня. Они пользовались обыкновенными, а также центральными разностями. Основополагающая работа Грегори (James Gregory) (1638—1675) была продолжена Ньютоном (1642—1727); Стирлинг (1692—1770) и позднее Бессель (1784—1846) добавили еще ряд формул. Строгая разработка теории интерполяции начинается с Гана (Hahn) и Фехера (Fejér, 1918). Опасности, связанные с равноотстоящим интерполированием полиномами, были обнаружены независимо друг от друга Рунге (1904) и Борелем (1903).

2. Интерполяция с помощью простых разностей

Одним из фундаментальных понятий анализа является ряд Тейлора, представляющий аналитическую функцию $f(x)$ через ее производные, взятые в точке $x=0$:

$$f(x) = f(0) + f'(0)x + f''(0)\frac{x^2}{2!} + \dots + f^{(n)}(0)\frac{x^n}{n!} + \dots \quad (5-2.1)$$

Эта формула может быть заменена аналогичной, содержащей конечные разности вместо производных. Пусть функция $f(x)$ задана в равноотстоящих точках

$$x = 0, h, 2h, 3h, \dots, nh, \dots \quad (5-2.2)$$

Образует последовательные «разностные коэффициенты»

I порядка:

$$\frac{\Delta f(0)}{\Delta x} = \frac{f(h) - f(0)}{h},$$

II порядка:

$$\frac{\Delta^2 f(0)}{\Delta x^2} = \frac{f(2h) - 2f(h) + f(0)}{h^2} \quad (5-2.3)$$

и т. д. Эти последовательные разностные коэффициенты можно считать разностными аналогами последовательных производных $f^{(k)}(0)$. Входящие в ряд Тейлора функции

$$\varphi_k(x) = \frac{x^k}{k!} \quad (5-2.4)$$

характеризуются функциональным уравнением

$$\varphi'_k(x) = \varphi_{k-1}(x). \quad (5-2.5)$$

При переходе к ряду с конечными разностями эти функции следует заменить функциями такого класса, которые удовлетворяют уравнению

$$\frac{\Delta \varphi_k(x)}{\Delta(x)} = \frac{\varphi_k(x+h) - \varphi_k(x)}{h} = \varphi_{k-1}(x). \quad (5-2.6)$$

Решение этого уравнения имеет вид

$$\varphi_k(x) = \frac{x(x-h)(x-2h)\dots(x-kh+h)}{k!}. \quad (5-2.7)$$

Наши формулы очень выиграют в простоте, если для независимой переменной x мы выберем такой масштаб, который делает h равным 1. Разностный коэффициент k -го порядка может быть тогда заменен просто разностью k -го порядка, а вспомогательные функции $\varphi_k(x)$ обращаются в функции

$$\varphi_k = \frac{x(x-1)(x-2)\dots(x-k+1)}{k!}, \quad (5-2.8)$$

которые и входят в *интерполяционную формулу Грегори — Ньютона*:

$$f(x) = f(0) + \Delta f(0)x + \Delta^2 f(0) \frac{x(x-1)}{2!} + \dots \\ \dots + \Delta^k f(0) \frac{x(x-1)\dots(x-k+1)}{k!} + \dots \quad (5-2.9)$$

Ряд Тейлора имеет характер «экстраполяционного» (предсказывающего) ряда, так как значение $f(x)$ и всех ее производных *в начальной точке* $x=0$ служит для вычисления $f(x)$ вне этой точки. Ряд (9), напротив, может служить для вычисления $f(x)$ *между* заданными точками 0, 1, 2, 3, ... Следовательно, он принадлежит к интерполяционному типу.

Практическое использование формулы (9) значительно облегчается, если составить «таблицу разностей» по следующему образцу:

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
0	10	11	— 5	3	— 4
1	21	6	— 2	— 1	
2	27	4	— 3		
3	31	1			
4	32				

(5-2.10)

Если, например, требуется найти значение $y=f(x)$ в точке $x=0,4$, то применение формулы (9) даст:

$$y(0,4) = 10 + 11 \cdot 0,4 - 5 \frac{(0,4)(-0,6)}{2} + 3 \frac{(0,4)(-0,6)(-1,6)}{6} - \\ - 4 \frac{(0,4)(-0,6)(-1,6)(-2,6)}{24} + \dots \cong 15,36.$$

Нулевая строка формулы (9) «подвижна» и может быть передвинута к любому целому значению переменной x . Если бы, например, требовалось найти $f(2,1)$, то мы, естественно, стали бы интерполировать, используя значения последовательных разностей в точке $x=2$ вместо $x=0$; тогда

$$f(2,1) = 27 + 4 \cdot 0,1 - 3 \frac{(0,1)(-0,9)}{2} + \dots \cong 27,54.$$

Мы выиграем, таким образом, в сходимости, но уменьшим число доступных членов, если спустимся далеко в низ схемы. Этот недочет можно, однако, исправить, заполняя пустые места внизу. Если, например, функция $f(x)$ мало отличается в рассматриваемом интервале от полинома четвертой степени, то в последнем столбце таблицы (10) на всех местах должны стоять числа -4 . Исходя из этого, можно последовательно заполнять пустые места по диагоналям (идущим снизу вверх). Так, например, новая дополнительная диагональ заполнится последовательно числами -4 , -5 , -8 , -7 , -25 (при движении справа налево). Дополненная таблица (10) примет вид

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$	
0	10	11	-5	3	-4	
1	21	6	-2	-1	-4	
2	27	4	-3	-5		
3	31	1	-8			
4	32	-7				
5	25					

(5-2.10')

Мы получаем теперь возможность использовать при вычислении $f(2,1)$ добавочное слагаемое

$$-5 \frac{(0,1)(-0,9)(-1,9)}{6} = -0,14.$$

3. Интерполяция при помощи центральных разностей

Простые разности целесообразно использовать лишь для начала и конца таблицы значений функции. («Начало» и «конец» суть понятия относительные и взаимно заменяются при обращении порядка следования значений x . Поэтому можно считать, что и случай «конца» таблицы уже рассмотрен в предыдущем параграфе.) Однако в случае среднего положения x в таблице мы можем с успехом использовать тот факт, что каждое значение функции имеет соседнее значение как справа, так и слева. Только в начале таблицы мы лишаемся соседних значений слева и в конце — соседних значений справа. Как только мы отходим от концов таблицы, можно построить таблицу «центральных разностей», сужающуюся от концов к середине в форме двух треугольников, симметричных относительно средней линии.

Принцип работы с таблицей центральных разностей также состоит в том, что мы вычисляем разность двух последовательных значений в одном и том же столбце. Но только мы по-иному используем эти разности. Разность записывается не рядом с верхним значением, а посередине между строкой верхнего и строкой нижнего значения. Поэтому мы различаем «полные строки» и «половинные строки». Исходные данные записываются в полных строках. Значения функции и разности располагаются по следующей схеме:

x	y	δy	$\delta^2 y$	$\delta^3 y$	$\delta^4 y$
0	10	11			
1	21	6	— 5	3	
2	27	4	— 2	— 1	— 4
3	31	1	— 3	— 5	— 4
4	32	— 7	— 8		— 3
5	25	— 23	— 16	— 8	
6	2				

(5-3.1)

В такой форме таблица полна пропусков. Как в полных строках, так и в половинных строках выписан лишь каждый второй член. Мы заполним теперь эти пропуски с помощью следующего простого приема, который относится ко *всем* столбцам: *берем среднее арифметическое числа, стоящего над пропуском, и числа, стоящего под пропуском*. Таким образом, дополненная схема таблицы (1) центральных разностей будет окончательно выглядеть следующим образом:

x	y	δy	$\delta^2 y$	$\delta^3 y$	$\delta^4 y$
0	10				
0,5	15,5	11			
1	21	8,5	— 5		
1,5	24	6	— 3,5	3	
2	27	5	— 2	1	— 4
2,5	29	4	— 2,5	— 1	— 4
3	31	2,5	— 3	— 3	— 4
3,5	31,5	1	— 5,5	— 5	— 3,5
4	32	— 3	— 8	— 6,5	— 3
4,5	23,5	— 7	— 12	— 8	
5	25	— 15	— 16		
5,5	13,5	— 23			
6	2				

(5-3.2)

Большое преимущество этой центрально-разностной схемы состоит в ее лучшей сходимости. Например, в строке 3 мы находим для второй и третьей разностей числа -3 , -3 , тогда как при первоначальном расположении эти значения были бы -8 , -8 . Кроме того, так как в нашем распоряжении, кроме целых строк, имеются еще половинные строки, то мы можем интерполировать от обоих типов строк, поэтому величина x в интерполяционной формуле может не превышать $\pm 0,25$, в то время как при применении прежней таблицы она не превышала $\pm 0,5$. Это тоже является большим преимуществом, так как увеличивает точность при том же числе разностей. Если пользоваться центральными разностями, то интерполяция с помощью трех последовательных разностей почти всегда будет достаточной при разумно распределенных исходных данных.

Фундаментальные функции $\varphi_k(x)$, соответствующие таблице центральных разностей, определяются функциональным уравнением

$$\varphi_k(x+1) + \varphi_k(x-1) - 2\varphi_k(x) = \varphi_{k-2}(x) \quad (5-3.3)$$

при начальных условиях $\varphi_0(x) = 1$, $\varphi_1(x) = x$. Удовлетворяющие этому уравнению функции четного и нечетного порядка определяются следующими формулами:

$$\left. \begin{aligned} \varphi_{2k}(x) &= \frac{x^2(x^2-1)(x^2-4)\dots[x^2-(k-1)^2]}{(2k)!}, \\ \varphi_{2k-1}(x) &= \frac{x(x^2-1)(x^2-4)\dots[x^2-(k-1)^2]}{(2k-1)!}. \end{aligned} \right\} \quad (5-3.4)$$

Функции для половинных строк были найдены Бесселем. Они входят в интерполяционные формулы Стирлинга и Бесселя. Формула Стирлинга, при которой полная строка является базисной, имеет следующий вид:

$$\begin{aligned} y(x) = & y_0 + (\delta y)_0 x + (\delta^2 y)_0 \frac{x^2}{2} + (\delta^3 y)_0 \frac{x(x^2-1)}{6} + \\ & + (\delta^4 y)_0 \frac{x^2(x^2-1)}{24} + (\delta^5 y)_0 \frac{x(x^2-1)(x^2-4)}{120} + \dots, \end{aligned} \quad (5-3.5)$$

а формула Бесселя, использующая в качестве базисной *половинную строку*, записывается в виде

$$\begin{aligned} y(x) = & y_0 + (\delta y)_0 x + (\delta^2 y)_0 \frac{x^2 - \frac{1}{4}}{2} + (\delta^3 y)_0 \frac{x \left(x^2 - \frac{1}{4} \right)}{6} + \\ & + (\delta^4 y)_0 \frac{\left(x^2 - \frac{1}{4} \right) \left(x^2 - \frac{9}{4} \right)}{24} + (\delta^5 y)_0 \frac{x \left(x^2 - \frac{1}{4} \right) \left(x^2 - \frac{9}{4} \right)}{120} + \dots \end{aligned} \quad (5-3.6)$$

В качестве примера рассмотрим вычисление при помощи этих формул двух значений функции y , определенной таблицей (2). Точка $x = 3,18$

близка к полной строке $x=3$, и поэтому при вычислении $y(3,18)$ целесообразно воспользоваться формулой Стирлинга. В строке 3 находим значения 31; 2,5; —3; —3; и, следовательно, получаем

$$y(3,18) = 31 + 2,5 \cdot 0,18 - 3 \frac{(0,18)^2}{2} - 3 \frac{(0,18)(0,18^2 - 1)}{6} = 31,49.$$

С другой стороны, допустим, что надо вычислить $y(3,41)$. Здесь точка лежит близ половинной строки 3,5, и мы можем воспользоваться формулой Бесселя при $x = -0,09$. Значения, стоящие в этой строке, суть 31,5; 1; —5,5; —5, и мы получаем

$$y(3,41) = 31,5 - 1 \cdot 0,09 - 5,5 \frac{(0,09)^2 - 0,25}{2} + 5 \cdot 0,09 \frac{(0,09)^2 - 0,25}{6} = 32,06.$$

В заключение рассмотрим вопрос об интерполяции в случае *неравноотстоящих значений аргумента*. Операции с «разделенными разностями»*) очень громоздки. Однако можно часто избежать их применения на основе следующих соображений. Введем вспомогательную переменную t с помощью некоторого соотношения

$$x = u(t). \quad (5-3.7)$$

Тогда табулированию по равноотстоящим значениям t будет соответствовать табулирование по неравноотстоящим значениям x . Но если изменение шкалы последнего табулирования достаточно гладко, то мы можем принять, что интерполяция производится по переменной t , а не по переменной x . Тогда прежние интерполяционные формулы сохраняют свою силу, несмотря на переменную шкалу. Единственное изменение состоит в том, что в предыдущих формулах надо заменить x через

$$t = \frac{x - x_0}{x_1 - x_0}, \quad (5-3.8)$$

где x_0 и x_1 — те две точки, между которыми мы интерполируем. Изложенные соображения показывают, что при табулировании функции $f(x)$ принятое ограничение строго равноотстоящими точками не является абсолютно решающим для применения интерполяционной формулы Стирлинга или Бесселя, если только изменение шкалы происходит по достаточно гладкому математическому закону¹⁾.

*) То есть с разностными коэффициентами вида $\frac{f(x+h) - f(x)}{h}$ при меняющемся значении h .

¹⁾ Об общем интерполировании полиномами в произвольных точках с помощью интерполяционной формулы Лагранжа см. VI, 10. Об остроумном методе Айткена (и его дальнейшем видоизменении Невиллем), который сводит интерполяцию полиномами к последовательным линейным интерполяциям, см. {4}, стр. 84; [4], стр. 76.

4. Дифференцирование табулированной функции

Операция $\frac{d}{dx}$ не может быть непосредственно применена к функции, которая задана лишь в определенных дискретных равноотстоящих точках. Однако после интерполирования функция $f(x)$ нам известна повсюду, и можно теперь произвести дифференцирование. Обычно нас интересует наклон кривой в тех же точках, в которых произведены наблюдения. При этом можно продифференцировать наши интерполяционные формулы при $x=0$ и получить таким образом соотношение между операциями $\frac{d}{dx}$ и $\frac{\Delta}{\Delta x}$.

Простые разности. В случае простых разностей, используя формулу Грегори—Ньютона (2.9), получим

$$f'(0) = \Delta f(0) - \frac{\Delta^2 f(0)}{2} + \frac{\Delta^3 f(0)}{3} - \dots \quad (5-4.1)$$

Это приводит к следующему операторному соотношению:

$$D = \Delta - \frac{\Delta^2}{2} + \frac{\Delta^3}{3} - \dots = \log(1 + \Delta). \quad (5-4.2)$$

Сходимость этой формулы очень медленная.

Центральные разности. Гораздо лучшая сходимость получается в случае центральных разностей. Дифференцирование формулы Стирлинга приводит к соотношению

$$D = \delta - \frac{\delta^3}{6} + \frac{\delta^5}{30} - \dots \quad (5-4.3)$$

Сходимость этой формулы намного лучше по сравнению с формулой (2) для случая простых разностей. Большим достоинством формулы (3) является также тот факт, что в нее входят только разности *нечетного* порядка. Еще более быстрая сходимость получается, если продифференцировать формулу Бесселя (3.6). Тогда получим

$$D = \delta - \frac{1}{24} \delta^3 + \frac{3}{640} \delta^5 - \dots \quad (5-4.4)$$

Множитель при δ^3 в 4 раза, а множитель при δ^5 в 7 раз меньше, чем на целых строках. Мы видим, таким образом, что наклон кривой может быть установлен с особенно большой точностью на *полпути между* точками наблюдения.

5. Неудобства таблицы разностей

Хотя исчисление конечных разностей является исключительно важным средством прикладного анализа, мы можем им пользоваться лишь с большой осторожностью, если имеем дело с эмпирическими функциями. Точность математически вычисленной таблицы какой-либо

функции обычно достаточна, чтобы составить по ней таблицу разностей. Если, однако, $y=f(x)$ просто *наблюдается* через равные интервалы, то разности высокого порядка обладают неприятным свойством значительно увеличивать ошибки наблюдения.

Допустим, что в ряду наблюдений наши данные точны, за исключением одного наблюдения, где оказалась ошибка на 1 единицу. Посмотрим, как эта единственная ошибка распространяется по всей схеме центральных разностей:

				0
			0	1
0	0	1	1	— 4
1	1	— 2	— 3	6
0	— 1	1	3	— 4
	0	0	— 1	1
			0	0

Из приведенной таблицы видно*), что единственный «выступ» расширяется все более и более, точно гора, возрастая и вширь и ввысь. Увеличение разностей происходит по биномиальному закону. Хотя это увеличение не кажется слишком быстрым, оно в действительности достаточно, чтобы во многих случаях свести на нет пользу таблицы разностей. Не надо забывать, что операция с разностями более высокого порядка предполагает определенную гладкость функции. Это значит, что по мере возрастания порядка разности быстро убывают и скоро становятся пренебрежимо малыми. Для функции, которую табулируют через малые интервалы — как, например, обыкновенные таблицы логарифмов, — редко имеют значение разности выше второго порядка. Здесь интерполяция сводится к одному-двум членам. Напротив, пользование разностями более высокого порядка может быть весьма полезным, делая возможным составление таблиц, рассчитанных через гораздо бóльшие интервалы. Но тогда необходимо, чтобы значения функции задавались с очень *большой точностью*. Таким образом, тесно расположенная таблица, вычисленная с небольшим числом десятичных знаков, может быть заменена таблицей, редко расположенной, при условии, что узловые значения заданы с точностью, достаточной для того, чтобы иметь возможность вычислять разности достаточно высокого порядка. Но такое условие, как правило, не выполняется в том случае, когда узловыми значе-

*) Здесь предположено, что все точные, т. е. не содержащие ошибок, значения табулированной функции равны нулю.

ниями таблицы являются величины, не вычисляемые математически, а наблюдаемые физически. Интервал, через который производятся наблюдения, обычно достаточно мал для того, чтобы сделать интерполяцию с двумя-тремя центральными разностями достаточно точной, однако трудность состоит в том, что увеличение ошибок в разностях высшего порядка полностью маскирует истинные значения этих разностей в связи с двумя противоположными влияниями: сильным уменьшением разностей и сильным увеличением ошибок. Поэтому разностная таблица эмпирически наблюдаемой функции обладает свойствами, весьма отличными от таблицы математически вычисленной функции. Первая и вторая разности не слишком отличаются от того, что должно быть на самом деле, *но относительные погрешности уже велики*. Разности более высокого порядка носят уже совсем случайный характер и не могут годиться для вычислений.

В силу этих обстоятельств при интерполировании эмпирических, а не математических функций следует исключить методы, использующие разности высокого порядка. В таких случаях необходимы другие методы эффективного интерполирования и дифференцирования эмпирических функций. Эти методы основаны на принципе «наименьших квадратов».

Одно частное свойство таблицы разностей заслуживает, однако, особого упоминания. Если среди наблюдений имеются *отдельные грубые ошибки*, совершенно разрушающие гладкий ход функции, то в случае применения пробной таблицы центральных разностей такие ошибки выступают точно ярко освещенные пятна на темном фоне. Четвертая или пятая разность обнаруживает в этом случае исключительно высокие максимумы, окруженные членами чередующегося знака. Такое поведение разностей локализует «дурные» места в наших измерениях, и следует сначала исправить исходные данные таким образом, чтобы аномальные пики исчезли. Это можно сделать, разделив пиковое значение на биномиальный коэффициент $(-1)^k C_{2k}^k$, если пик оказался в $2k$ -й разности, и вычитая это число из входного в той строке, в которой находится пик.

6. Основной принцип метода наименьших квадратов

Предположим, что две переменные связаны функциональной зависимостью, математический закон которой известен гипотетически, но некоторые параметры этого закона остаются неизвестными. Если имеется столько исходных измерений, сколько неизвестных параметров, то определение этих последних есть чисто алгебраическая задача *). Если, однако, мы произвели больше наблюдений, чем это необходимо, то мы имеем систему уравнений

*) Точнее, задача, требующая решения системы, число уравнений которой равно числу неизвестных.

которую можно назвать «переопределенной», ибо число уравнений больше числа неизвестных. Если бы измерения были свободны от ошибок, то мы могли бы просто отбросить излишние уравнения, ибо они не добавляют ничего нового к прежним. Так как, однако, измерения неточны, то каждое новое наблюдение добавляет некоторую новую информацию. Уравнения в том виде, как они получились, вообще говоря, несовместны и, следовательно, не могут быть решены; их надо «подогнать». Для этого мы берем разность между «теоретическими» значениями и теми, которые получены из наблюдения, называя эти разности «невязками». Хотя невозможно, вообще говоря, найти такие значения неизвестных параметров, которые делали бы каждую из этих невязок равной нулю, мы всегда можем найти такие значения этих параметров, которые делают *сумму квадратов невязок минимальной*. Величина этого минимума дает меру точности наших наблюдений. В предположении, что постулированный нами математический закон точен, нулевой минимум означал бы, что каждая невязка равна нулю, и следовательно, все наши наблюдения совершенны. Чем больше этот минимум, тем более ошибочны наши наблюдения. Квадратный корень из суммы квадратов минимальных невязок можно считать «средней ошибкой» наших измерений, хотя иногда целесообразно проверить *распределение* этих невязок, чтобы увидеть, не окажутся ли некоторые невязки подозрительными по своей величине. Если обнаружатся невязки, более чем втрое превышающие среднюю ошибку, то следует исключить соответствующие им измерения и привести к минимуму сумму квадратов остальных невязок.

Во многих задачах физики и техники неизвестные параметры входят в данный математический закон *линейно*. В этом случае задача приведения к минимуму суммы квадратов невязок эквивалентна приведению к минимуму алгебраического выражения второй степени. Получающиеся при этом уравнения линейны, а матрица их коэффициентов симметрична и неотрицательна (в большинстве хорошо сформулированных задач эта матрица даже положительно определена). Эти уравнения называются «нормальными уравнениями» данной задачи.

Тот же принцип применим ко всем задачам разложения по системам ортогональных функций. Задача здесь состоит в том, чтобы разложить заданную функцию в линейную комбинацию заранее выбранных функций. Если эта последовательность конечна, то мы не можем получить точного ответа. Все же мы можем всегда получить «наилучший» ответ, беря разность между заданной функцией и ее разложением и рассматривая эту разность как невязку нашего приближения. Так как нас интересует аппроксимирование функций не только в одной какой-нибудь точке, а всюду внутри данной непрерывной области, то метод наименьших квадратов теперь требует проинтегрировать квадрат невязки по данной области. Получающиеся нормальные уравнения дают тогда «наилучшее»

приближение функции $f(x)$ в этой области, выраженное через выбранные функции. Этот прием является основой теории систем ортогональных функций и имеет фундаментальное значение почти во всех разделах прикладного анализа.

7. Сглаживание эмпирических данных при помощи четвертых разностей

Задача интерполирования ряда наблюдений, в которых сказывается разброс, вызванный случайными ошибками, отличается от соответствующей задачи для математически табулированной функции. Заданные точки обычно столь близки друг к другу, что простая линейная интерполяция была бы вполне достаточна. Однако затруднение состоит в том, что сами основные значения неточны и дали бы весьма «ухабистую» кривую, если бы мы просто соединили их отрезками прямых. Наши данные необходимо «сгладить», чтобы вывести из них дальнейшие заключения. Это значит, что на основе некоторых статистических соображений следует уменьшить влияние случайных ошибок. Один простой метод, являющийся аналитическим аналогом построения кривой сглаживания, основан на использовании четвертых разностей.

Будем рассуждать следующим образом. Примем, что наши данные достаточно близки друг к другу, чтобы оправдать предположение о том, что вторая производная не изменяется существенно на протяжении нескольких измерений. Комбинируем каждое измерение с двумя соседними измерениями слева и справа. Это дает нам совокупность пяти последовательных измерений. Если бы отсутствовали ошибки наблюдений, то соответствующие точки на графике располагались бы около параболы второго порядка. При этих условиях ход измерений дается, следовательно, законом

$$y = a + bx + cx^2, \quad (5-7.1)$$

где коэффициенты a , b , c должны быть подобраны соответственно нашим данным. Однако эти три параметра не могут быть подогнаны к пяти данным, поэтому мы воспользуемся принципом наименьших квадратов. Образует разность

$$\sum_{k=-2}^{+2} (y - y_k)^2 \quad (5-7.2)$$

и приводим к минимуму эту величину надлежащим выбором a , b , c . В нашем случае данные соответствуют точкам $x = -2, -1, 0, 1, 2$, и следовательно, функция, которую нужно сделать минимальной,

есть

$$\begin{aligned} & (a - 2b + 4c - y_{-2})^2 + \\ & + (a - b + c - y_{-1})^2 + \\ & + (a - y_0)^2 + \\ & + (a + b + c - y_1)^2 + \\ & + (a + 2b + 4c - y_2)^2. \end{aligned} \quad (5-7.3)$$

Условие минимальности дает для a и c два «нормальных уравнения»:

$$\left. \begin{aligned} 5a + 10c - \sum_{k=-2}^{+2} y_k &= 0, \\ 10a + 34c - \sum_{k=-2}^{+2} k^2 y_k &= 0. \end{aligned} \right\} \quad (5-7.4)$$

(Уравнение для b не представляет для нас интереса в настоящий момент, ибо нашей целью является исправление центрального значения y_0 , которое соответствует значению $x=0$. Но, как показывает уравнение (1), значение функции при $x=0$ есть $y=a$. Следовательно, нам нужно лишь a .) Решение системы (4) относительно a дает

$$\begin{aligned} 70a &= -6y_{-2} + 24y_{-1} + 34y_0 + 24y_1 - 6y_2 = \\ &= 70y_0 - 6(y_{-2} - 4y_{-1} + 6y_0 - 4y_1 + y_2) = 70y_0 - 6\delta^4 y_0, \end{aligned} \quad (5-7.5)$$

где $\delta^4 y_0$ обозначает четвертую центральную разность в нулевой строке. Мы находим, таким образом, что

$$a = y_0 - \frac{3}{35} \delta^4 y_0 \quad (5-7.6)$$

и получаем следующий простой метод, с помощью которого часто может быть эффективно уменьшен «случайный разброс» достаточно плотного ряда данных. Строим таблицу центральных разностей до четвертого порядка включительно. Затем вычитаем из каждой ординаты $\frac{3}{35}$ четвертой разности, соответствующей ей в той же строке (так как $\frac{3}{35}$ очень близко к $\frac{1}{12}$, то можно считать, что практически поправка составляет $-\frac{1}{12} \delta^4 y_0$). Каждая отдельная ордината может быть исправлена указанным образом. Полученные новые данные будут соответствовать более гладкой кривой. В этом можно убедиться, построив снова таблицу разностей и увидев, что четвертые разности теперь значительно уменьшены и менее изменчивы, чем они были раньше. Часто бывает целесообразно брать в исправленных данных на один десятичный знак больше, чем в первоначальных данных. Это обеспечит более гладкий ход. Если четвертые разности

снова обнаруживают значительные неправильности в ходе, то этот метод может быть повторен еще раз ¹⁾).

Следующий численный пример иллюстрирует этот метод в приложении к измерениям, полученным при взлетном испытании самолета. Наблюдения производились через каждую секунду, но в таблице приведено лишь каждое *второе* измерение. Это сделано потому, что измерения разбиты были на две группы—измерения четных и нечетных секунд. Каждая группа измерений была независимо от другой подправлена по одному и тому же методу, и полученные результаты были сравнены. Это доставило ценную проверку результатов, дав возможность оценить, в какой мере разброс данных измерения обусловлен флуктуацией. Нижеследующая таблица содержит расчеты только с одной группой измерений (см. таблицу на стр. 326). Другая группа была обработана идентично и дала совпадающие результаты.

Первая и последняя пары исходных данных (приведенных в таблице) не охватываются общим приемом, так как мы здесь не имеем соседних значений с обеих сторон. Мы вынуждены, следовательно, пользоваться в этом случае ближайшими значениями с *одной только стороны* и поэтому прокладывать параболу второй степени $a + bx + cx^2$ через пять последовательных точек. Для первой пары исходных данных дело сводится к той же процедуре, что и раньше, ибо мы можем условиться, что значение $x = 0$ соответствует третьей точке, а первые две точки обозначить через $x = -2$ и $x = -1$. Теперь, однако, нам нужны теоретические значения (1) также и для этих двух точек, в предположении, что

$$\tilde{y}_{-2} = a - 2b + 4c \quad \text{и} \quad \tilde{y}_{-1} = a - b + c$$

взяты в качестве исправленных значений первых двух измерений.

¹⁾ Если наблюдения настолько часты, что 7 точек, а именно 3 соседних точки, с обеих сторон центральной точки и она сама, могут быть соединены параболой второго порядка, то поправочная формула принимает вид

$$\bar{y}_0 = y_0 - \frac{9\delta^4 y_0 + 2\delta^6 y_0}{21}.$$

В обоих случаях можно избежать операции с таблицей центральных разностей, воспользовавшись «техникой подвижной полосы» типа (8.5). Кодовые числа подвижной полосы в случае двух соседних точек будут

$$\frac{-3, 12, \overset{\downarrow}{17}, 12, -3}{35},$$

а в случае трех соседних точек —

$$\frac{-2, 3, 6, \overset{\downarrow}{7}, 6, 3, -2}{21}.$$

Стрелка указывает «нулевую строку», в которую следует вписать результирующую поправку.

t	y	δy	$\delta^2 y$	$\delta^3 y$	$\delta^4 y$	$-\frac{3}{35} \delta^4 y$	$\bar{y}$
0	0						0,8
2	25	25,0					20,7
4	67,4	42,4	17,4				75,6
6	172	104,6	62,2	44,8	-95,5	8,2	170,1
8	288,1	116,1	11,5	-50,7	22,1	-1,9	276,6
10	387,1	187,7	-17,1	-28,6	134,4	-11,5	404,3
12	574,8	99,0	88,7	105,8	-201,2	17,2	563,9
14	755,8	206,1	-6,7	95,4	127,2	-10,9	759,0
16	961,9	225,9	31,8	31,8	37,1	3,2	960,4
18	1187,8	257,8	25,1	-5,3	17,4	-1,5	1190,2
20	1445,6	273,7	19,8	12,1	-28,1	2,4	1441,6
22	1719,3	320,1	31,9	-16,0	46,5	-4,0	1724,4
24	2039,4	337,2	46,4	30,5	-59,8	5,1	2038,8
26	2376,6	332,5	17,1	-29,3	7,5	-0,6	2373,9
28	2709,1	337,2	-4,7	-21,8	31,2	-2,7	2710,3
30	3046,3	337,1	4,7	-4,8	-14,2	1,2	3046,4
32	3383,4		-0,1				3383,3

Вычисления дают следующий результат:

$$\begin{aligned}\tilde{y}_{-2} &= y_{-2} + \frac{1}{5} \delta^3 + \frac{3}{35} \delta^4, \\ \tilde{y}_{-1} &= y_{-1} - \frac{2}{5} \delta^3 - \frac{1}{7} \delta^4,\end{aligned}$$

если считать при этом, что для δ^3 и δ^4 мы берем ближайшие вычисленные значения соответствующих центральных разностей, находящихся по верхней наклонной линии таблицы разностей. Например, поправка для $y(0)$, т. е. y_{-2} , дает в нашем случае

$$\tilde{y}(0) = 0 + \frac{1}{5} (44,8) + \frac{3}{35} (-95,5) = 0,8,$$

а поправка для $y(2)$, т. е. y_{-1} , будет

$$\bar{y}(2) = 25 - \frac{2}{5}(44,8) - \frac{1}{7}(-95,5) = 20,7.$$

Соответствующие формулы для другого конца таблицы:

$$\begin{aligned}\tilde{y}_2 &= y_2 - \frac{1}{5}\delta^3 + \frac{3}{35}\delta^4, \\ \tilde{y}_1 &= y_1 + \frac{2}{5}\delta^3 - \frac{1}{7}\delta^4,\end{aligned}$$

если и здесь взять ближайшие соответствующие разности, расположенные вдоль нижней наклонной линии таблицы разностей. Для нашего примера эти формулы дают

$$\begin{aligned}\tilde{y}(32) &= 3383,4 - \frac{1}{5}(-4,8) + \frac{3}{35}(-14,2) = 3383,2, \\ \tilde{y}(30) &= 3046,3 + \frac{2}{5}(-4,8) - \frac{1}{7}(-14,2) = 3046,4.\end{aligned}$$

8. Дифференцирование эмпирической функции

Определение производной как предела разностного коэффициента мало что дает, если наши наблюдения не свободны от ошибок. Отношение $\frac{\Delta y}{\Delta x}$ становится очень чувствительным даже к небольшим ошибкам, когда Δx делается весьма малым. Поэтому мы не можем в таких случаях применять те методы дифференцирования равноотстоящих данных, какими пользовались для математически точных данных. При решении задачи дифференцирования вновь возникает необходимость в методе наименьших квадратов.

Предположение, сделанное в предыдущем параграфе, а именно, что вторая производная мало изменяется на протяжении пяти последовательных наблюдений, выполняется во многих случаях. Мы можем поэтому провести параболу второй степени через эти точки, но парабола должна быть найдена при помощи метода наименьших квадратов, ибо пять точек, вообще говоря, не будут лежать на параболе второй степени. Процедура точно такая же, как и та, которой мы следовали в § 7. Мы комбинируем каждую точку с соседними точками по две с каждой стороны. Опять приводим к минимуму сумму (7.3) и снова получаем нормальные уравнения (7.4). Единственная разница в том, что в прежней задаче нужно было получить исправленное значение функции в точке $x=0$, тогда как теперь нужно получить исправленное значение *производной* в той же точке.

Процедура для двух первых и последних ординат. В начале и конце наших наблюдений мы лишены части соседних значений, так как располагаем ими только по одну сторону. Это нарушает симметрию процесса интерполирования и значительно снижает точность. Поэтому мы должны провести параболу через первые несколько точек и использовать производную в точках $t=0$ и $t=1$. В начале кривой целесообразно, по-видимому, выбрать лишь четыре точки вместо пяти, ибо здесь физические условия еще недостаточно установились, и мы не можем рассчитывать на гладкость кривой в такой же мере, как в дальнейшем.

Решение нормальных уравнений для данной задачи дает следующую формулу для первых двух недостающих скоростей:

$$\left. \begin{aligned} f'(0) &= \frac{-21f(0) + 13f(h) + 17f(2h) - 9f(3h)}{20h}, \\ f'(h) &= \frac{-11f(0) + 3f(h) + 7f(2h) + f(3h)}{20h}. \end{aligned} \right\} \quad (5-8.6)$$

В нашем численном примере первые две недостающие скорости поэтому будут равны

$$\begin{aligned} y'(0) &= \frac{27}{20} = 1,35, \\ y'(1) &= \frac{237}{20} = 11,85. \end{aligned}$$

Формулы (6) применимы также для конца ряда наблюдений; в этом случае ординаты $f(0)$, $f(h)$, $f(2h)$, $f(3h)$ должны быть заменены значениями $f(x_n)$, $f(x_n - h)$, $f(x_n - 2h)$, $f(x_n - 3h)$, и формулы (6) дают $-f'(x_n)$ и $-f'(x_n - h)$.

Может случиться, что ряд наших исходных данных расположен не настолько тесно, чтобы оправдать предположение, что пять последовательных точек лежат практически на параболе второго порядка. В таком случае может оказаться более надежным объединить лишь *четыре* точки для одного сглаживания. Мы сохраним симметрию, если условимся, что скорости будут получаться в точке, расположенной *посредине между данными*. Подвижная полоса теперь содержит только четыре числа: -3 , -1 , 1 , 3 , а нулевая строка лежит посредине между ∓ 1 . Теперь не хватает только по *одному* значению в начале и в конце таблицы. Это значение получается теперь как среднее арифметическое двух значений производных (6):

$$f'\left(\frac{h}{2}\right) = \frac{-8f(0) + 4f(h) + 6f(2h) - 2f(3h)}{10}. \quad (5-8.7)$$

Применяя этот второй метод к прежнему ряду наблюдений, мы

заданной функции. Формулы (4.3) и (4.4) показывают, что для вычисления производной нужно пользоваться центральными разностями нечетного порядка. Но эти разности оказываются практически негодными вначале ввиду наличия помех, и притом в тем большей степени, чем меньше h , так как разности уменьшаются вместе с h , тогда как помехи остаются на том же уровне, создавая тем самым недопустимо большую ошибку. С другой стороны, таблица разностей может быть распространена *влево* в виде разностей *отрицательного* порядка, что означает в действительности *суммирование* вместо вычитания. Вредное влияние помех в этих случаях значительно уменьшается ввиду случайного характера ошибок, которые в процессе суммирования имеют тенденцию сглаживаться.

Следующая формула интегрирования по частям непосредственно показывает, что процесс, выраженный формулой (8.4), может быть проведен, если мы имеем в своем распоряжении два первых левых столбца таблицы разностей:

$$\int_b^a x u' dx = |xu|_a^b - \int_a^b u dx. \quad (5-9.4)$$

Если положить $u' = y$ и обозначить u через $D^{-1}y$, а интеграл от u через $D^{-2}y$, то

$$\int_b^a xy dx = |xD^{-1}|_a^b - |D^{-2}|_a^b. \quad (5-9.5)$$

Мы покажем теперь, что таблица (8.5) может быть получена численным приемом, который бывает иногда даже более быстрым, чем прежний алгоритм. Образует для этого два столбца разностей отрицательного порядка, т. е. сумм, а затем сумм этих сумм первичных данных. В целях упрощения расположения мы запишем эти столбцы справа от столбца данных величин, хотя с точки зрения структуры таблицы разностей они должны стоять слева:

x	y	Σy	$\Sigma^2 y$
0	0	0	0
1	4	4	4
2	25	29	33
3	50	79	112
4	67,4	146,4	258,4
5	124,9	271,3	529,7
6	172,0	443,3	973,0
7	201,4	644,7	1617,7
8	288,1	932,8	2550,5
9	321,3	1254,1	3804,6
10	387,1	1641,2	5445,8

(5-9.6)

Рассмотрим строку $x=8$, $y=288,1$. Так как в (8.5) были использованы две соседние точки, то мы доходим до строки $x=10$, а число в нижней строке $x=6$ уменьшается на 1, давая, таким образом, $x=5$. В столбце Σy находим два числа 1641,2 и 271,3. Применяя формулу (5), нужно x в первом члене правой части положить равным 2 в верхнем пределе, а в нижнем пределе $x=-2$; поэтому образуем сумму $1641,2 + 271,3$ и умножаем ее на 2:

$$(1641,2 + 271,3) \cdot 2 = 3825,0.$$

Теперь переходим ко второму члену формулы, который находится в столбце $\Sigma^2 y$. Здесь мы доходим до строк $10 - 1 = 9$ и $6 - 1 = 5$ и находим там 3804 и 529,7, разность которых составляет

$$3804,6 - 529,7 = 3274,9.$$

Теперь, имея оба члена, находим их разность.

$$3825,0 - 3274,9 = 550,1.$$

Это значение совпадает, если не считать положения запятой, со значением $y'(x)$, имеющимся в таблице (8.5) против входного значения $y=288,1$. Таким образом, мы получили другой метод построения скорости изменения эмпирических данных.

Во многих задачах техники нас интересует интеграл от функции, а также *момент* интеграла. Метод, изложенный в этом параграфе, показывает, как эти величины могут быть получены построением столбцов Σy и $\Sigma^2 y$.

10. Вторая производная эмпирической функции

Затруднения, с которыми мы встретились при устранении помех в задаче дифференцирования, еще значительнее в том случае, когда надо найти *вторую* производную от эмпирической функции. Такие задачи часто возникают при анализе данных, полученных при наблюдении движения. Действительно, в аэродинамических задачах приходится получать выводы, касающиеся действия сил. Необходимо, следовательно, из измерений перемещений находить ускорения. Эти трудности можно иллюстрировать тем фактом, что даже резкое изменение в силе вызовет лишь небольшое отклонение в перемещении. Сглаживание исходных данных обуславливает небольшую погрешность, если речь идет о самих перемещениях, но вторая производная может быть изменена при этом процессе весьма значительно. Поэтому мы не можем ожидать большой точности при численном подсчете второй производной.

Вместо того, чтобы стараться получить непосредственно вторую производную, лучше действовать в два этапа; сначала получить (изложенным ранее методом) первую производную, а затем применить тот же процесс к первой производной. Так как определение произ-

водной в каждой отдельной точке охватывает пять первичных наблюдений, а определение одной точки второй производной включает пять точек первой производной, то, как легко видеть, *девять последовательных точек* первоначальной кривой определяют одну точку кривой ускорения. Весьма вероятно, что такой процесс приводит к выравниванию случайных ошибок, но он действует неблагоприятно, если кривая ускорений в действительности не является гладкой в силу более или менее резких изменений в силе.

Если нас не интересует кривая скоростей и мы хотим получить непосредственно кривую ускорений, то обе указанные операции можно объединить в одну операцию с подвижной полосой. Учитывая, что для образования второй производной мы умножаем пять значений первой производной на «кодовые числа» — 2, — 1, 0, 1, 2, соответственно, а при образовании каждого из этих значений первой производной мы аналогично действовали на пять первоначальных данных, нетрудно подсчитать, что при процессе непосредственного образования второй производной из основных данных кодовые числа для девяти последовательных данных будут

$$\frac{4, 4, 1, -4, -10, -4, 1, 4, 4}{100h^2}.$$

(5-10.1)

Этот процесс вычисления очень удобен, так как все умножения на 4 могут быть объединены в одну операцию. Мы снова можем нанести кодовые числа на подвижную полосу и заставить скользить эту полосу вдоль столбца данных. Результат будет соответствовать второй производной к моменту, противолежащему кодовому числу — 10.

В качестве численного примера мы выполним этот процесс для данных таблицы (8.5), причем получим теперь вторую производную $y''(x)$ через каждую секунду. В нашей задаче $h=1$, и деление, указанное в знаменателе формулы (1), сводится просто к перенесению запятой на 2 знака влево:

	y	y''
	0	
	4	
	25	
	50	
	67,4	7,97
	124,9	5,98
	172	4,64
	201,4	...
	288,1	
	321,3	
	387,1	
	...	

(5-10.2)

Процедура для первых четырех и последних четырех ординат. Так как в общем процессе используется по четыре соседних значения с обеих сторон каждого наблюдения, то мы теряем первые четыре значения ускорения в начале таблицы, и то же повторяется в конце таблицы. Однако мы можем восстановить эти значения, хотя и с меньшей надежностью, пользуясь отдельно техникой формулы (8.6) для первых (и последних) двух ординат кривой скоростей. Окончательный результат может быть выражен в виде весов, которые должны быть приписаны первым данным наблюдений. Мы получаем отдельные ряды весовых множителей для первой, второй, третьей и четвертой точек кривой ускорения. Они представлены в последовательных столбцах следующей таблицы:

	$f''(0)$	$f''(h)$	$f''(2h)$	$f''(3h)$
y_0	115	85	53	26
y_1	— 116	— 76	— 33	— 4
y_2	— 124	— 84	— 51	— 18
y_3	118	58	13	— 14
y_4	25	15	2	— 8
y_5	— 18	2	8	2
y_6			8	8
y_7				8

Результат разделить на $200 h^2$

(5-10.3)

В нашем численном примере (2) применение этих весов приводит к следующим результатам:

$$\left. \begin{aligned} y''(0) &= 8,86, \\ y''(1) &= 8,78, \\ y''(2) &= 8,76, \\ y''(3) &= 7,66. \end{aligned} \right\}$$

(5-10.4)

Мы уже отмечали при рассмотрении процесса нахождения скорости, что комбинирование пяти последовательных данных может оказаться неприемлемым, если промежутки между наблюдениями недостаточно малы. Поэтому таблица (8.8) была рассчитана на основе только четырех соседних наблюдений, и мы получили скорости для точек, расположенных посередине между двумя данными точками. После повторения этого же процесса первоначальное положение точек восстанавливается. Теперь используются только *три* вместо четырех соседних точек с каждой стороны. Кодовые числа для кривой ускорений теперь будут

$$\frac{9, 6, -5, -20, -5, 6, 9}{100},$$

(5-10.5)

и мы теряем только три вместо четырех ординат в начале и в конце кривой.

Так как первая производная, которую мы получаем, есть $f' \left(\frac{h}{2} \right)$, а производные этих данных начинаются с $f''(h)$, то мы не можем указать $f''(0)$. Следовательно, мы должны оставить свободным место против первого и последнего наблюдения. Веса для дальнейших двух ускорений даны в следующей таблице, заменяющей более обширную таблицу (3):

	$f''(h)$	$f''(2h)$
y_0	52	27
y_1	— 54	— 14
y_2	— 44	— 29
y_3	36	1
y_4	16	6
y_5	— 6	9

(5-10.6)

Результат разделить
на $100 h^2$

Сравнение изложенной техники получения ускорений с помощью трех и четырех соседних значений для случая измерений при взлетных испытаниях дается в нижеследующей таблице. Она содержит первые 20 ускорений для данных, полученных при взлетных испытаниях, вычисленных один раз на основе четырех смежных $\langle y'' \rangle$, и один раз на основе трех смежных значений (y'') с обеих сторон каждого наблюдения:

y	$\langle y'' \rangle$	(y'')	y	$\langle y'' \rangle$	(y'')
0	8,86		387,1	8,22	9,90
4	8,78	8,13	500,6	3,60	4,01
25	8,76	7,97	574,8	3,39	— 1,59
50	7,86	8,59	650,1	3,81	3,84
67,4	7,97	8,08	755,8	6,26	7,46
124,9	8,98	6,31	815,8	10,17	11,70
172	4,64	4,03	961,9	8,25	8,31
201,4	6,12	4,39	1051,8	7,49	8,52
288,1	6,58	6,09	1187,8	5,21	3,91
321,2	8,76	11,33	1321,6	5,49	3,27

(5-10.7)

Ход кривой ускорений в общем сходен при обоих вычислениях. Однако амплитуды флуктуаций гораздо больше во втором случае, когда учитывалось три вместо четырех смежных данных. Из расхождения этих двух серий результатов можно вывести заключение, что наблюдения через каждую секунду при этих взлетных испытаниях не являются достаточно близкими друг к другу, чтобы можно было провести удовлетворительное сглаживание, не нарушая в то же время истинный ход кривой. Ход самолета в первой фазе не гладок и далек от тех установившихся условий, которые господствуют на более

позднем этапе полета. Неустановившиеся колебания, вызванные сочетанием горизонтального и вертикального движения, дают себя чувствовать даже в то время, когда самолет находится еще на стартовой дорожке. Условие, что ускорение не меняется существенно на протяжении пяти или даже четырех измерений, не было выполнено в нашей задаче. Необходимо было бы иметь четыре-пять отсчетов в секунду в качестве базы нашего анализа для того, чтобы получить надежную кривую ускорений, в которой исключены случайные ошибки измерений и в то же время существенно сохранен истинный ход ускорения.

11. Сглаживание в целом с помощью разложения в ряд Фурье

В наших предыдущих рассуждениях каждое наблюдение комбинировалось со своими непосредственными соседями слева и справа. Мы пользовались аналитическим характером $f(x)$ в окрестности каждой точки и старались исключить неаналитическое поведение помех с помощью таких операций, которые не выходят за пределы непосредственной окрестности данной точки x . Мы можем, таким образом, говорить о «сглаживании в малом», или «локальном» приеме исключения помех. Теперь мы рассмотрим совершенно другой подход. Вместо того, чтобы разбивать наши наблюдения на ряд окрестностей, мы можем рассмотреть *всю совокупность исходных данных* как единое целое и постараться найти средства отделить истинный ход функции от налагающихся помех. Этот метод сглаживания имеет то преимущество, что он лишь в небольшой степени зависит от каких-либо специальных предположений относительно природы неизвестной функции $f(x)$. В наших прежних соображениях мы предполагали, например, что в некоторой конечной окрестности точки вторая производная функции $f(x)$ практически постоянна. Физически это значит, что сила, действующая на движущееся тело, лишь слабо меняется в пределах выбранного промежутка времени. Такие предположения не всегда могут выполняться в действительных аэродинамических условиях. Резкий порыв ветра, например, вызывает внезапное и непредвиденное изменение второй производной, которое не удовлетворяет нашей гипотезе непрерывности. Локальный прием сглаживания имеет тенденцию сглаживать эти разрывы и тем самым изменять истинный ход второй производной, исходя из предположения равномерности, которое не соответствует действительной физической картине. Поэтому желателен метод, который не был бы ограничен такими предположениями.

Всякий раз, когда рассматриваются задачи по сглаживанию функции в целом, на сцену в качестве одного из самых мощных математических средств выступают ряды Фурье. Поэтому можно попытаться исследовать задачу помех с помощью ряда Фурье.

Мы уже раньше отмечали, что помеха не обладает свойствами гладкости и дифференцируемости, присущими аналитическим функциям. Это обстоятельство заставляет нас отнести помехи в специальную категорию функций по отношению к рядам Фурье. Строго говоря, ряд Фурье есть бесконечный ряд, но его сходимость дает возможность ограничиться при исследовании первыми n членами ряда. Ответ на вопрос, насколько большим следует выбрать это n , полностью определяется аналитическими свойствами функции, которая разлагается в ряд. Если функция всюду непрерывна, но ее производная претерпевает разрыв в некоторой точке, то члены ряда Фурье убывают со скоростью n^{-2} . Если сама функция становится разрывной в некоторой точке, то члены ряда убывают лишь со скоростью n^{-1} . Кратковременный импульс подвергает опасности сходимость ряда, а формальное разложение бесконечно острого импульса («дельта-функции») действительно оказывается расходящимся. Но «помеху» можно понимать как нерегулярное следование острых импульсов, и поэтому гармоническое разложение помехи вообще не будет сходиться. Здесь, следовательно, открывается возможность различить истинный ход функции от наложенной на нее помехи. Гармоническое разложение функции обнаружит *более быструю сходимость*, чем гармоническое разложение помехи.

Для того чтобы воспользоваться этим характерным отличием в сходимости между функцией и помехой, необходимо сделать его возможно более резким. Пусть задано большое число наблюдений в точках

$$x = 0, h, 2h, \dots, nh = l. \quad (5-11.1)$$

Если не принять надлежащих мер предосторожности, то ряд Фурье, составленный для аппроксимирования неизвестной функции $f(x)$, будет очень слабо сходиться даже в том случае, когда $f(x)$ всюду ведет себя нормально, но не выполняются условия на границе. Если ни $f(x)$, ни $f'(x)$ при $x = l$ не принимают значений, которые они имели при $x = 0$, то разрыв на границе будет определять характер сходимости. Мы можем улучшить сходимость, отразив $f(x)$ как четную функцию для отрицательных x , и таким образом разложить $f(x)$ в ряд по одним косинусам. Это избавит нас от разрывности самой функции, но *производная* все же будет разрывной на границе. Однако мы можем пойти еще дальше. Вычтем из $f(x)$ надлежаще выбранный двучлен $\alpha + \beta x$, рассматривая, таким образом, функцию

$$g(x) = f(x) - (\alpha + \beta x). \quad (5-11.2)$$

Коэффициенты α и β определим из условий на границе:

$$g(0) = 0, \quad g(l) = 0. \quad (5-11.3)$$

Затем отразим $g(x)$ для отрицательных x как нечетную функцию

$$g(-x) = -g(x). \quad (5-11.4)$$

Если такую функцию сделать периодической с периодом $2l$, то ни она, ни ее первая производная не будут иметь разрывов. Разрыв появится только во второй производной. Асимптотический порядок величины членов ряда Фурье будет теперь n^{-3} , а такая сходимость практически является удовлетворительной.

В соответствии со сказанным разложим функцию $g(x)$ в ряд синусов:

$$g(x) = b_1 \sin \frac{\pi}{l} x + b_2 \sin \frac{2\pi}{l} x + \dots \quad (5-11.5)$$

Так как мы располагаем значениями

$$y_k = f(kh) \quad (k = 0, 1, 2, \dots, n), \quad (5-11.6)$$

то сначала совершаем преобразование

$$g(x) = f(x) - f(0) - \frac{f(l) - f(0)}{l} x \quad (5-11.7)$$

(этим мы добиваемся выполнения условий (3)). Затем определяем коэффициенты b_k разложения (5) из условия, что в заданных точках $x = kh$ ряд должен давать видоизмененные основные данные $g(kh)$, т. е. первоначальные измерения, скорректированные двучленом $\alpha + \beta x$. Это условие дает

$$b_k = \frac{2}{n} \sum_{\alpha=1}^{n-1} g(\alpha h) \sin k\alpha \frac{\pi}{n}. \quad (5-11.8)$$

Но процесс сглаживания всегда базируется на том факте, что в нашем распоряжении имеется гораздо больше исходных измерений, чем это необходимо для сглаживания функции. Гармоническое разложение функции $f(x)$ не содержит обертонов выше определенной частоты, обычно называемой «граничной частотой» ν_0 . Это означает, что свыше некоторого заранее определяемого индекса коэффициенты Фурье b_k практически равны нулю. Наивысший порядок $k = m$, который еще должен рассматриваться, определяется соотношением

$$\frac{m\pi}{l} x = 2\pi\nu_0 x, \quad (5-11.9)$$

или

$$m = 2\nu_0 l = 2\nu_0 nh. \quad (5-11.10)$$

Это дает условие

$$\frac{m}{n} = 2\nu_0 h. \quad (5-11.11)$$

Произведение $2\nu_0 h$, т. е. удвоенная граничная частота, умноженная на промежуток времени между двумя измерениями, есть отвлеченное число, определяющее эффективность, с которой мы можем сгладить наши измерения. Число, обратное ему, мы обозначим через p и назовем «параметром сглаживания»:

$$p = \frac{1}{2\nu_0 h}. \quad (5-11.12)$$

Если p меньше 1, то это означает, что наши исходные измерения настолько разрознены, что они не могут определить ход кривой $f(x)$ даже при отсутствии помех. Если $p=1$, то мы имеем как раз минимальное число наблюдений, необходимых для определения $f(x)$, и ничего не остается для сглаживания. Таким образом, p должно быть больше 1 для того, чтобы можно было вообще произвести сглаживание. Вообще, параметр сглаживания p означает *отношение фактического числа исходных измерений к их минимальному числу, которое требуется при отсутствии помех*. Эффективное сглаживание требует, чтобы p равнялось по меньшей мере 2, но мы едва ли получим удовлетворительные результаты, если оно менее 4—5.

Допустим сначала, что из физических соображений мы можем заранее установить положение граничной частоты ν_0 . Тогда можно рассуждать следующим образом. Если бы помехи отсутствовали, то коэффициенты Фурье $b_1, b_2, \dots, b_m$ имели бы определенные значения, следующие же коэффициенты равнялись бы практически нулю. Тогда разложение Фурье

$$g(x) = b_1 \sin \frac{\pi}{l} x + b_2 \sin \frac{2\pi}{l} x + \dots + b_m \sin \frac{m\pi}{l} x \quad (5-11.13)$$

интерполировало бы надлежащим образом нашу функцию не только в заданных точках, но и во *всех* остальных точках интервала.

Наличие помех меняет картину. Коэффициенты $b_{m+1}, \dots, b_{n-1}$ теперь уже не равны нулю и не обнаруживают тенденции уменьшаться. Они представляют несходящуюся часть ряда Фурье, обусловленную неаналитическим характером помех. Идеально «случайные» помехи имели бы спектр Фурье, в котором нет предпочтения какой-нибудь частоте и который поэтому давал бы среднюю амплитуду для всех частот, со случайными флуктуациями. Если наше разложение содержит синусы и косинусы, то мы могли бы говорить «о случайном распределении фазы», в то время как амплитуды сохраняли бы одинаковый постоянный порядок величины. Так как наше разложение содержит лишь синусы, то случайность фазы заменяется случайным следованием знаков в распределении амплитуд $b_k (k > m)$. Учтем теперь, что помехи влияют на амплитуды b_k разложения двояким образом. Амплитуды, расположенные за b_m , полностью обусловлены помехами. В этой части спектра мы можем исключить помехи,

просто опустив в ряде Фурье все члены, следующие за $k = m$. Первоначально мы приняли, что ряд Фурье имеет вид

$$g(x) = \sum_{k=1}^{n-1} b_k \sin k \frac{\pi}{l} x, \quad (5-11.14)$$

и определили коэффициенты b_k из условия, что наши данные наблюдений должны быть полностью представлены; теперь мы обрываем ряд, образуя сумму

$$\bar{g}(x) = \sum_{k=1}^m b_k \sin k \frac{\pi}{l} x. \quad (5-11.15)$$

Этот процесс исключает все высокочастотные составляющие помехи. Правда, компоненты b_k при $k < m$ также подвергались в некоторой степени влиянию помехи. Однако в этой части спектра мы лишены возможности отделить помеху от истинных компонент, так как случайное распределение знаков мешает нам вычесть компоненту, соответствующую помехам. Эта неопределенность неизбежна. Все же нам удалось исключить большую часть помех благодаря тому, что мы отбросили ту часть ряда Фурье, которая полностью обусловлена ими.

На практике мы редко знаем частоту ν_0 заранее. Даже тогда, когда известно из физических соображений, что измеренная функция $g(x)$ не может содержать частоты выше некоторой определенной ν_0 , все же нет никакой гарантии, что *все* частоты вплоть до ν_0 являются подлинными. Гладкий характер функции $g(x)$ может быть таков, что подлинный спектр ее заканчивается гораздо раньше, чем ν_0 , и все амплитуды после этой частоты являются ложными и должны быть отброшены. Чем меньше число компонент подлинной функции $\bar{g}(x)$, тем более успешно мы исключаем помехи наших измерений.

Вычитание линейного двучлена $\alpha + \beta x$ из наших исходных наблюдений дало возможность свести негладкость на границе к разрыву не раньше чем во второй производной. Разрыв второй производной часто встречается в действительной физической практике. В задачах движения разрывность второй производной означает резкое изменение силы. Такое изменение, математически вызванное нашим желанием сделать функцию периодической, может иметь место при измерениях также и физически, вследствие, например, резкого порыва ветра, исчерпания горючего, изменения режима работы двигателя и подобных обстоятельств. Следовательно, физические причины негладкости задачи имеют такой же характер, что и негладкость на границе, и закон убывания коэффициентов n^{-3} не может быть улучшен, даже если бы функция была по природе периодической.

12. Эмпирическое определение граничной частоты ν_0

Мы поступим гораздо разумнее, если не будем пытаться заранее определить ν_0 , а найдем решение этого вопроса непосредственно в самих имеющихся данных. Убывание коэффициентов Фурье по закону n^{-3} является достаточно сильным для вполне эффективного разделения аналитической и неаналитической частей спектра. Мы производим для наших исходных данных полный анализ Фурье, прикидываем величину компонент b_k , определяемых как ординаты, соответствующие значениям абсциссы $k = 1, 2, 3, \dots$. Сначала величины b_k оказываются большими и не обнаруживают какой-либо заметной закономерности в своем расположении. Но затем они довольно круто уменьшаются до относительно малых значений, которые далее не убывают; мы замечаем, что теперь амплитуды остаются в пределах некоторой полосы плюс-минус значений. Они не обнаруживают тенденции становиться значительно большими или много меньшими некоторой средней величины. Эту среднюю величину можно установить, если, начиная с конца спектра, подсчитать сумму

$$\beta^2 = \frac{1}{N} (b_{n-1}^2 + b_{n-2}^2 + \dots + b_{n-N}^2). \quad (5-12.1)$$

При N возрастающем β будет приближаться к почти постоянному значению. Тогда мы проводим горизонтальные прямые $y = \pm \beta$ и определяем точку $k = m$, в которой эти прямые пересекают низкочастотную часть спектра. Мы сохраняем все b_k вплоть до $k = m$ и опускаем все b_k для $k > m$. *Это простое усечение ряда Фурье первыми m членами осуществляет сглаживание наших данных.*

Невозможно точно определить ту точку $k = m$, на которой мы должны оборвать ряд. Неизбежна некоторая «неясная зона», в которой коэффициенты b_k , обусловленные функцией, и коэффициенты, обусловленные помехами, имеют один и тот же порядок величины. Однако эта неопределенность не имеет решающего значения. Неизбежная неопределенность, обусловленная низкочастотными помехами, которые нельзя исключить, содержится и в аналитических компонентах. При наличии этой неопределенности несущественно, добавили ли мы несколько ошибочных коэффициентов Фурье или, наоборот, отбросили несколько принадлежащих искомой функции коэффициентов, которые мы приписали помехам. Важно только, чтобы эта неясная зона не была слишком обширной. Но такой опасности не существует, ибо закон n^{-3} обеспечивает настолько быстрое убывание подлинных амплитуд, что их можно отделить от неубывающих компонент помех.

Следующий численный пример служит для иллюстрации этого метода. Последовательность 68 данных летных испытаний была разложена в ряд Фурье по синусам. После вычитания линейного дву-члена $\alpha + \beta x$ коэффициенты Фурье были вычислены по формуле (11.8)

(при $n=67$) с помощью цифровой вычислительной машины (IBM). Анализ дал следующие значения для 66 компонент функции $g(x)$:

k	$-b_k$	k	$-b_k$	k	$-b_k$	k	$-b_k$
1	2024,40	17	2,33	33	4,50	49	1,27
2	200,65	18	-5,05	34	-2,47	50	-3,54
3	115,14	19	2,65	35	2,99	51	1,83
4	20,35	20	-1,80	36	-2,85	52	-0,23
5	16,08	21	2,53	37	-0,48	53	2,83
6	5,62	22	-0,87	38	-0,07	54	1,71
7	8,31	23	-1,58	39	0,02	55	-3,59
8	6,89	24	0,60	40	2,29	56	2,50
9	0,95	25	-4,33	41	-1,14	57	-4,35
10	7,62	26	4,26	42	0,49	58	2,93
11	-2,30	27	-3,82	43	-2,42	59	0,34
12	3,24	28	4,70	44	-0,46	60	-0,12
13	1,97	29	-4,45	45	-0,04	61	0,62
14	1,31	30	0,55	46	0,58	62	-2,57
15	3,89	31	-1,13	47	2,18	63	0,12
16	-2,86	32	-2,14	48	-1,91	64	-1,19
						65	2,25
						66	0,67

Анализ этой таблицы показывает, что помехи в настоящей задаче имеют не только случайный характер. Последовательность знаков обнаруживает определенную закономерность со многими систематическими чередованиями $+$, $-$. Такие закономерности могут встретиться, если из наблюдений не исключены «грубые ошибки», которые выходят за пределы среднего порядка величин ошибок. Мы могли бы исключить эти «плохие места» с самого начала с помощью вычитания; это сделало бы помехи менее согласованными. Однако метод рядов Фурье имеет то преимущество, что эти плохие места не портят существенно главную часть функции, а просто увеличивают до некоторой степени помехи. Поэтому мы получаем удовлетворительные результаты, не удаляя из исходных данных явно плохие наблюдения.

В нашей задаче «неясная зона» невелика. Ордината $b_{10}=7,62$ определенно лежит вне полосы помехи и должна быть включена в аналитическую часть функции. Мы могли бы закончить ряд на ней, но можно также включить в него члены b_{11} и b_{12} и закончить ряд на $k=12$. Неопределенность, таким образом, не имеет существенного значения. Мы останавливаемся на $k=12$ и опускаем все гармоники, следующие за двенадцатой. Число наблюдений, следовательно, почти в 5,5 раза больше, чем число нужных компонент. Мы можем поэтому принять, что в усеченном ряде, кроме того, что он сглажен, исключено почти 80% всех помех. Остальные 20%

присутствуют в виде искажения оставшихся первых двенадцати коэффициентов b_k .

Проблема дифференцирования ряда еще остается неразрешенной. Тот факт, что для нахождения перемещения можно пренебречь всеми гармониками сверх двенадцати, не означает, что это же будет верно для скорости и, еще в меньшей мере для ускорения. Из того, что мы получили хорошее приближение для $f(x)$, не следует, что мы имеем хорошее приближение для $f'(x)$ или для $f''(x)$. Производная от хорошего приближения функции не обязательно будет хорошим приближением для производной от функции. Функция $f''(x)$ может потребовать для своего представления значительно большего числа членов ряда Фурье, чем $f(x)$. Но эти члены более высокого порядка не могут быть получены путем дифференцирования ряда Фурье для $f(x)$ ввиду того, что коэффициенты b_k , соответствующие этим значениям k , почти полностью обусловлены помехами и не показательны для поведения $f(x)$ при отсутствии помех.

При этих условиях нам остается вернуться к локальным операциям, с помощью которых может быть определена производная. По самой своей природе производная определяется значениями функции $f(x)$ в окрестности данной точки. Закон «произведение массы на ускорение равно движущей силе» показывает, что в достаточно малом промежутке времени ускорение не может изменяться слишком быстро. Мы не сильно ошибемся, если допустим, что пять последовательных наблюдений лежат на параболе третьего порядка. Тогда вторая производная имеет еще достаточно свободы для линейного изменения в течение этого промежутка времени. Если это условие не выполняется, то наши наблюдения вообще слишком разрознены, чтобы допускать эффективное сглаживание. Кубическая парабола $y = \alpha + \beta x + \gamma x^2 + \delta x^3$ имеет четыре степени свободы, и следовательно, мы определяем четыре константы на основе пяти измерений. Таким образом, степень переопределенности невелика, и мы не нарушаем требований точности нашей задачи. Если мы возьмем вторую производную от этого полинома третьей степени (полученного методом наименьших квадратов) в точке симметрии, то получим следующие симметрические веса пяти последовательных ординат взамен прежних весовых множителей формулы (10.1):

$$\frac{2, -1, -2, -1, 2}{7h^2}. \quad (5-12.3)$$

В прежней схеме каждое наблюдение комбинировалось с *четырьмя* соседними с каждой стороны. Всего, таким образом, девять последовательных данных определяли одну точку кривой ускорений. Это оправдано, если наблюдения следуют друг за другом через достаточно малые интервалы. Но если это условие не выполняется, то прежний метод (10.1) приведет к чрезмерному сглаживанию.

Мы потеряем некоторые детали кривой ускорения, которые совершенно реальны и не вызваны помехами. Необходимо иметь в виду, что задача сглаживания имеет два аспекта. С одной стороны мы должны по возможности исключить помеху; с другой стороны, мы не должны исключать деталей, которые на самом деле относятся к наблюдаемой функции. У нас нет оснований для того, чтобы заранее принять, что кривая ускорений будет особенно гладкой, если речь идет о полете самолета. Резкие порывы ветра могут нарушать действие регулярных аэродинамических сил; но даже нормальные аэродинамические силы сами по себе могут вызвать сложные отклонения от установившегося полета. Если мы хотим реалистически изучить кривую ускорений, то мы постараемся избежать излишнего сглаживания кривой. Использование весовых множителей по схеме (3) имеет лучшие шансы на точность, чем использование весов по схеме (10.1), так как оно включает лишь пять, вместо девяти, последовательных данных. С другой стороны, теперь возникает опасность, что нам не удалось достаточно полно исключить помехи. Полученная при этом кривая ускорений может оказаться слишком негладкой из-за ошибок наблюдений.

Здесь можно привлечь для дополнительного сглаживания приемы гармонического анализа. Если n есть общее число наблюдений, и мы вычислили $n - 1$ коэффициентов ряда синусов для гармонического анализа данных ускорения, полученных с помощью локальных квадратичных парабол, то мы можем принять, что не все эти коэффициенты действительно будут нужны для представления истинного ускорения. Наивысшие коэффициенты соответствуют высокочастотным колебаниям, которые по аэродинамическим соображениям следует считать весьма неправдоподобными. Поэтому более раннее усечение ряда надлежащим образом сгладит кривую ускорений, не нарушая существенно требований надежности. Таким образом, сочетание легкого локального сглаживания с дополнительным сглаживанием путем усечения ряда Фурье можно рассматривать как наиболее правдоподобное решение задачи сглаживания, если требуется большая осторожность в связи с тем, что имеющиеся данные не следуют друг за другом столь близко, чтобы можно было произвести эффективное сглаживание при помощи только локальных парабол.

Численный пример на стр. 345 иллюстрирует влияние двух разных приемов сглаживания. Из ряда 68 наблюдений при взлете самолета были получены значения ускорения в каждой точке кривой. Столбец s содержит данные фактического смещения в футах через промежутки времени в $h = 0,96$ секунды. Столбец γ содержит ускорение, вычисленное по методу § 10. Первые четыре и последние четыре значения γ были получены с помощью таблицы (10.3).

Столбец a был составлен иначе. Веса были взяты согласно схеме (12.3) сначала без учета знаменателя $7h^2$. Первые два и последние два значения ускорения были получены в предположении, что локаль-

ная парабола, построенная в точке $k=2$ (и аналогично в точке $k=n-2$), может быть использована для получения ускорения в недостающих точках. Это дает следующую таблицу в соответствии с прежней таблицей (10.3):

	$f''(0)$	$f''(h)$	
y_0	9	5,5	(5-12.4)
y_1	-15	-8	
y_2	-2	-2	
y_3	13	6	
y_4	-5	-1,5	

Результаты
разделить на $7h^2$

k	s	γ	a	$\bar{a}$	$\bar{\gamma}$	k	s	γ	a	$\bar{a}$	$\bar{\gamma}$
0	117	-0,74	14	12,1	1,86	34	4042	2,85	19	16,3	2,51
1	144	2,38	22	17,7	2,72	35	4230	2,84	17	21,5	3,31
2	163	5,69	30	38,8	5,96	36	4425	2,57	23	21,9	3,37
3	192	8,34	74	67,1	10,31	37	4619	2,29	12	13,9	2,14
4	229	9,48	75	78,5	12,07	38	4819	1,82	12	6,0	0,92
5	281	8,70	67	62,9	9,67	39	5017	2,03	7	8,4	1,29
6	340	6,98	34	38,0	5,84	40	5218	2,52	11	18,5	2,84
7	407	5,55	26	26,8	4,12	41	5420	3,18	31	25,3	3,89
8	472	5,05	34	31,1	4,78	42	5623	3,48	25	23,7	3,64
9	545	5,11	39	35,8	5,50	43	5839	3,27	24	19,7	3,03
10	625	5,40	27	34,0	5,23	44	6047	3,53	5	20,5	3,15
11	706	6,02	35	34,8	5,35	45	6266	3,39	36	24,4	3,75
12	792	6,69	51	44,1	6,78	46	6479	3,51	28	25,5	3,92
13	887	7,01	49	52,7	8,10	47	6708	2,87	18	21,3	3,27
14	989	6,82	47	49,2	7,56	48	6933	1,95	5	13,7	2,11
15	1096	6,37	39	38,0	5,84	49	7157	1,37	16	5,5	0,84
16	1212	5,91	35	33,0	5,07	50	7389	0,90	3	3,6	0,55
17	1329	5,79	37	37,5	5,76	51	7618	1,68	- 6	4,1	0,64
18	1435	5,66	41	41,3	6,35	52	7845	2,79	19	18,1	2,77
19	1584	5,33	35	36,9	5,67	53	8075	3,97	43	32,9	5,05
20	1718	5,24	32	30,5	4,69	54	8312	4,72	30	36,3	5,58
21	1858	5,10	32	30,8	4,73	55	8557	4,67	24	29,5	4,53
22	2002	5,04	36	35,5	5,46	56	8798	4,57	32	25,8	3,97
23	2150	4,91	32	35,5	5,46	57	9049	5,00	37	30,7	4,72
24	2306	4,51	30	29,1	4,47	58	9305	3,63	17	30,0	4,61
25	2462	4,17	25	23,6	3,63	59	9562	-0,07	12	8,5	1,31
26	2625	3,82	28	23,9	3,67	60	9821	-3,30	-16	-22,8	-3,51
27	2790	3,69	18	26,0	4,00	61	10082	-4,55	-35	-33,2	-5,10
28	2959	3,71	27	25,4	3,90	62	10330	-0,32	-18	- 9,4	-1,44
29	3129	3,76	28	23,7	3,64	63	10578	2,04	32	25,3	3,89
30	3307	3,46	25	24,0	3,69	64	10830	2,00	42	38,4	5,90
31	3486	3,17	17	23,8	3,66	65	11092	0,99	18	25,2	3,87
32	3668	3,10	23	19,9	3,06	66	11356	0,64	11	7,7	1,18
33	3853	3,12	16	15,2	2,34	67	11616	0,30	4	0	0

Полученные таким образом значения (без общего знаменателя $7h^2$) содержатся в столбце a . Эти значения обнаруживают значительный разброс и требуют дополнительного сглаживания. Для этой цели был использован метод гармонического анализа. Применен был ряд Фурье из синусов в соответствии с рассуждениями § 11. В рассматриваемом случае удалось обойтись без вычитания линейного двучлена $\alpha + \beta x$,

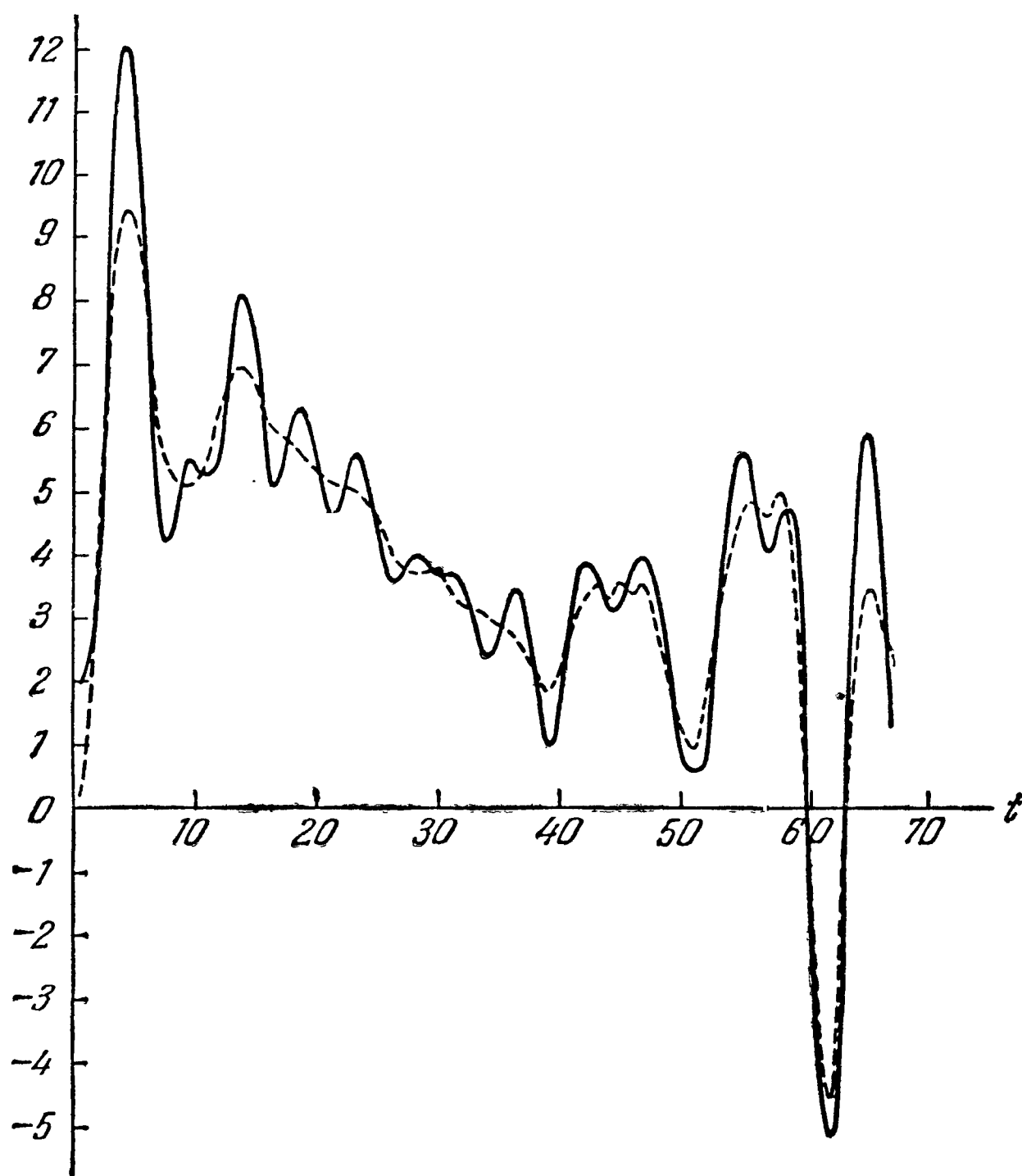


Рис. 34.

так как ускорение легко можно было экстраполировать до нуля в обоих концах ряда, и полученные данные были, таким образом, непосредственно пригодны для разложения в ряд синусов. Затем ряд был усечен до 30 членов. Сумма этих 30 членов, вычисленная для точек наблюдения, дала столбец $\bar{a}$. Наконец, путем деления чисел этого столбца на $7h^2$ был получен последний столбец $\bar{\gamma}$.

Для сравнения были вычерчены обе кривые γ и $\bar{\gamma}$. Как видно из рис. 34 (γ — пунктирная линия, $\bar{\gamma}$ — сплошная линия), девятиточечное сглаживание излишне выравняло кривую, стерев некоторые детали кривой ускорения, которые принадлежат самим исходным данным и не являются результатом процедуры сгла-

живания. Происхождение этих колебаний не может быть установлено на основании одних только исходных данных. Потребовалось бы тщательное исследование физических условий для того, чтобы получить больше информации относительно отдельных деталей кривой $\bar{\gamma}$. Гармонический анализ, несомненно, указывает на наличие некоторой высокочастотной компоненты. Но одни только исходные данные не дают основания решить, относятся ли эти колебания к аэродинамическим условиям или к измерительным приборам.

Приведенный пример иллюстрирует те трудности, с которыми мы можем встретиться при обработке данных, полученных при не оптимальных условиях. Если аэродинамический анализ не интересует нас во всех случайных подробностях явления взлета, то более сильное сглаживание кривой γ будет вполне достаточным для того, чтобы представить существенные черты явления. Но наличие дополнительных колебаний малых амплитуд может представлять при правильном понимании значительный интерес. Адекватное изучение этих более тонких особенностей требует использования измерительной аппаратуры более высокой точности и достаточно частого следования наблюдений. Тогда инструментальную часть помехи можно было бы исключить с помощью локальных парабол по методу наименьших квадратов. В этом случае после получения ряда данных для ускорения, которые в значительной мере свободны от инструментальных помех, мы можем подвергнуть эти данные дополнительному анализу Фурье. Компоненты, превышающие уровень помех, укажут на существование скрытых периодичностей, изучение которых может содействовать более глубокому пониманию аэродинамических явлений.

13. Полиномы метода наименьших квадратов

В задаче сглаживания мы встретились с процессом проведения параболы через четыре или пять заданных точек методом наименьших квадратов. Подобные ситуации встречаются во многих других задачах физики и техники; поэтому целесообразен общий разбор этого приема. Гаусс, первый рассматривавший эту задачу, ввел изящное обозначение, с помощью которого результирующие уравнения записываются в особенно ясной форме. Он использовал символ $[u]$ в следующем смысле. Величина u вычисляется во всех заданных точках, и все эти значения суммируются. Следовательно, если заданные точки отмечаются индексом $i = 1, 2, \dots, n$, то выражение $[u]$ обозначает

$$[u] = u_1 + u_2 + \dots + u_n; \quad (5-13.1)$$

поэтому $[x^k]$ будет означать $[x^k] = x_1^k + x_2^k + \dots + x_n^k$ и аналогично

$$[yx^k] = y_1 x_1^k + y_2 x_2^k + \dots + y_n x_n^k.$$

При этом не предполагается, что наблюдения сделаны обязательно для равноотстоящих значений независимого переменного x .

Общая задача интерполяции полиномом по методу наименьших квадратов может быть поставлена следующим образом. Имеются теоретические основания предположить, что ряд измерений

$$y_1, y_2, \dots, y_n, \quad (5-13.2)$$

неизвестной функции $y=f(x)$ при некоторых заранее выбранных значениях независимой переменной

$$x=x_1, x_2, \dots, x_n \quad (5-13.3)$$

может быть хорошо представлен полиномом m -й степени

$$y=a_0 + a_1x + \dots + a_mx^m. \quad (5-13.4)$$

Нам известна степень m полинома, тогда как коэффициентами $a_0, a_1, \dots, a_m$ мы можем распорядиться свободно и они должны быть определены на основе измерений. Мы принимаем, что число измерений превышает m . Если бы их было меньше, то задача не имела бы, вообще говоря, однозначного решения.

Неизвестные коэффициенты a_i полинома (13.4) определяются теперь с помощью следующего принципа. Для каждой точки наблюдения мы образуем «невязку»

$$a_0 + a_1x_i + \dots + a_mx_i^m - y_i \quad (5-13.5)$$

и берем сумму квадратов всех невязок;

$$Q = \sum_{i=1}^m (a_0 + a_1x_i + \dots + a_mx_i^m - y_i)^2. \quad (5-13.6)$$

Величина Q по своему характеру неотрицательна, причем нулевое значение возможно лишь в том случае, когда каждая невязка в отдельности равна нулю, т.е. если наши измерения все согласуются и в *точности* соответствуют полиному m -й степени. На это рассчитывать нельзя. Однако мы можем найти такие значения для a_i , при которых сумма Q становится *минимальной*. Мы считаем полином, соответствующий этим a_i , наилучшим образом представляющим наши измерения (ср. также § 16).

Эта задача нахождения минимума имеет однозначное решение: оно получается в результате решения некоторой системы линейных уравнений (с ненулевым определителем), которую мы сейчас выведем.

Согласно теореме дифференциального исчисления необходимое условие минимума требует, чтобы частные производные от Q по

и Бесселя, использующие центральные разности и дающие лучшую сходимость. При этих видах интерполяции мы считаем установленным, что функция может быть интерполирована с помощью степенного ряда. Хотя это предположение кажется весьма разумным, его справедливость в действительности ни в коем случае не гарантирована. Допустим, что имеется функция $f(x)$, определенная в бесконечном интервале от 0 до ∞ и даже аналитическая повсюду в этом интервале. Пусть заданы значения этой функции в бесконечно большом числе точек $x = 0, 1, 2, \dots$. Можем ли мы интерполировать эту функцию между заданными точками с помощью формул Грегори—Ньютона? Будет ли интерполяционный ряд сходиться и будет ли давать правильный ответ, если он сходится? Более глубокое исследование показывает, что только вполне определенный класс функций, определяемый некоторым интегральным преобразованием, допускает полиномиальное интерполирование типа Грегори—Ньютона. Функции, не принадлежащие к этому классу, дают расходящееся разложение, а это означает, что интерполяционная формула лишается смысла. Поэтому при интерполировании степенными рядами требуется большая осторожность.

Трудность обычно коренится в том факте, что при табулировании функций разности по мере возрастания порядка убывают так быстро, что через несколько этапов они исчезают. Это, однако, не означает, что интерполяционная формула, если взять достаточно большое число членов, уменьшит ошибку до произвольно малой величины. Во многих случаях может случиться нечто совершенно неожиданное. Ошибка становится все меньше и меньше благодаря корректирующему влиянию более высоких разностей, но затем достигается минимум, после чего погрешность снова увеличивается и становится произвольно большой. (Мы имеем в виду математически заданные функции, которые принципиально могут быть вычислены с любой точностью. В эмпирических функциях помехи исключают возможность использовать разности высоких порядков.)

Как можно объяснить это загадочное явление? Заметим, что центральные разности, стоящие вдоль определенной строки разностной таблицы, определяются исключительно соседними значениями, *непосредственно примыкающими* к данной точке в начале строки. По мере перехода к разностям более высокого порядка это «соседство» расширяется все более и более. Поэтому использование небольшого числа членов интерполяционной формулы отличается от использования большого числа членов следующим образом. В одном случае мы предполагаем пригодность полиномиального приближения *в малом*, в другом же случае приближение считается действительным *в целом*. Между тем, имеет место тот факт, что приближение полиномами в достаточно малой окрестности точки всегда надежно и оправдано, тогда как в целом оно не всегда надежно и требует надлежащей осторожности.

Вейерштрасс доказал в 1885 г., что всякая непрерывная функция, определенная в конечном интервале, всегда может быть с любой степенью точности аппроксимирована полиномами. Эта теорема устанавливает, следовательно, правомерность полиномиального разложения, и, казалось бы, мы не можем совершить ошибки, интерполируя наши данные степенным рядом. В действительности, однако, из теоремы Вейерштрасса, устанавливающей правильность приближения полиномами, не вытекает, что такое приближение может быть получено путем подгонки к равноотстоящим данным. Э. Борель в 1903 г. и О. Рунге в 1901 г. открыли тот поразительный факт, что можно взять такие очень простые аналитические функции, как

$$y = \frac{1}{1 + 25x^2} \quad (5-14.1)$$

в интервале $[-1, +1]$, и получить совершенно неправильные результаты при равноотстоящем интерполировании. Когда мы располагаем наши заданные точки все теснее и теснее друг к другу, интерполирующий полином, который содержит все наши точки, действительно неограниченно приближается к заданной функции $f(x)$ в большей части заданного интервала. Но после некоторой точки, которую можно заранее определить (в примере (1) — точка $x = \pm 0,726$), вплоть до конца интервала интерполирующий полином не стремится ни к какому пределу и фактически возрастает бесконечно в каждой точке этой части интервала. Так как Рунге исследовал это явление очень подробно, было бы правильно назвать его «явлением Рунге».

Итак, имеет место тот странный факт, что полином, который не содержит заданных точек, может оказаться в лучшем положении относительно точности в целом, чем полином, содержащий эти точки. Интересен, к примеру, следующий факт. Пусть $f(x)$ удовлетворяет условию на концах: $f(\pm 1) = 0$. Будем аппроксимировать ее полиномом степени $4n$. Но вместо того, чтобы расположить на нем $4n + 1$ равноотстоящих точек, подберем его так, чтобы на нем расположилось только $2n + 1$ узловых точек:

$$f\left(\frac{k}{n}\right) \quad (k = 0, \pm 1, \dots, \pm n).$$

Значения, соответствующие точкам оси абсцисс, расположенным посредине между абсциссами узловых точек, вычислим по следующей интерполяционной формуле:

$$f\left(\frac{k + \frac{1}{2}}{n}\right) = \frac{1}{\pi} \sum_{\alpha = -n}^{+n} (-1)^{\alpha - k} \frac{f\left(\frac{\alpha}{n}\right)}{\alpha - k + \frac{1}{2}}. \quad (5-14.2)$$

Следовательно, мы заменим точные значения функции в этих средних точках неточными значениями. Тем не менее, образованный таким

образом интерполирующий полином дает небольшие ошибки во всем интервале, тогда как «правильно» составленный интерполирующий полином дает ошибки неограниченно большие.

Затруднения, исследованные Рунге, обусловлены только равноотстоящим характером заданных значений. Если данные располагаются не эквидистантно, а распределены по нулевым точкам полинома Чебышева $(2n+1)$ -го порядка $T_{2n+1}(x)$, то затруднения исчезают. Ошибки интерполяции тогда колеблются вокруг величины одного и того же порядка на протяжении всего интервала и в каждой его точке стремятся к нулю при неограниченном возрастании числа исходных точек.

Так как равноотстоящие данные гораздо более удобны как с точки зрения вычислений, так и с точки зрения наблюдений, то можно поставить вопрос, как получить эффективные приближения полиномами, несмотря на равноотстоящий характер заданных точек. Дело в том, что во многих задачах физики и техники бывает особенно желательным заменить некоторую аналитическую функцию степенным рядом. Степенные функции переменной x обладают большими операционными преимуществами, и может появиться надобность воспользоваться ими даже в том случае, когда первоначальная функция $f(x)$ (табличная или полученная из наблюдений) не является степенным рядом. Мы знаем из теоремы Вейерштрасса, что такая замена всегда возможна. Но мы знаем также из исследований Рунге, что нельзя получить этот полином простым интерполированием. Эта задача не может быть также решена способом наименьших квадратов, так как мы не знаем заранее, какова будет степень аппроксимирующего полинома и желательно ли сводить к минимуму невязки, ибо малые невязки в заданных точках могут привести к большим погрешностям в промежуточных точках.

Вернемся снова к разложениям в ряд Фурье (§ 11), при помощи которого мы старались уменьшить помехи в исходных данных. Мы вычли подходящий линейный двучлен $\alpha + \beta x$ из заданных значений и затем использовали разложение в ряд по синусам. Далее мы отделили аналитическую часть функции от части, соответствующей помехам, путем исследования последних коэффициентов использованной части ряда Фурье и обрыва его в надлежащем месте. Мы таким образом получили аналитическое выражение, которое не только сглаживало наши заданные значения, но и интерполировало значения функции $f(x)$ между заданными точками. Этот метод тригонометрической интерполяции свободен от недостатков равноотстоящего интерполирования полиномами. Тригонометрические функции обладают свойствами ортогональности относительно равноотстоящих данных и поэтому занимают такое же преимущественное положение по отношению к этим данным, какое имеют степени x по отношению к данным, распределенным по нулевым точкам чебышевских полиномов. Тригонометрическая интерполяция с помощью ряда синусов будет автоматически сходиться к $f(x)$.

в каждой точке интервала, если исходные заданные точки будут располагаться все плотнее и плотнее.

Мы стремимся, однако, получить приближение функции $f(x)$ *полиномами*. Для этой цели мы теперь превратим наше разложение по синусам в полиномиальное разложение. Разложение Тейлора для каждого отдельного синуса в ряд Фурье и собирание подобных членов нецелесообразно. Такая процедура не привела бы к цели, так как полученный ряд слабо сходил бы и потребовалось бы поэтому большое число членов. Можно заранее сказать, что лучше всего воспользоваться для наших целей полиномами Чебышева, так как этот ряд даст наиболее быструю сходимость (ср. VII, 6). Поэтому нам надо исследовать выражение функции синус в виде ряда, расположенного по полиномам Чебышева.

В разложение (11.13), соответствующее исходным данным, нормированным к интервалу $[0, 2]$, входят функции

$$\varphi_k(x_1) = \sin k \frac{\pi}{2} x_1. \quad (5-14.3)$$

Удобнее, однако, поместить начало отсчета посредине интервала, отделив, таким образом, заранее четную и нечетную части нашей функции. Это означает, что x_1 следует заменить через $x = x_1 - 1$. Функции $\varphi_k(x)$ теперь принимают вид

$$\begin{aligned} \varphi_{2k}(x) &= (-1)^k \sin k\pi x, \\ \varphi_{2k+1}(x) &= (-1)^{k+1} \cos \left(k + \frac{1}{2} \right) \pi x. \end{aligned} \quad (5-14.4)$$

Первая группа функций дает нечетную, а вторая группа — четную часть функции $g(x)$; наконец, прибавление поправочного двучлена $\alpha + \beta x$ восстанавливает первоначальную функцию $f(x)$.

Разложение тригонометрических функций (4) по полиномам Чебышева достигается с помощью функций Бесселя $J_k(x)$. При этом получаются следующие результаты:

$$\begin{aligned} \sin k\pi x &= 2 \sum_{\alpha=1}^{\infty} (-1)^{\alpha} J_{2\alpha+1}(k\pi) T_{2\alpha+1}(x), \\ \cos \left(k + \frac{1}{2} \right) \pi x &= 2 \sum_{\alpha=0}^{\infty} ' (-1)^{\alpha} J_{2\alpha} \left(\left(k + \frac{1}{2} \right) \pi \right) T_{2\alpha}(x). \end{aligned} \quad (5-14.5)$$

Знак $\sum'$ во второй формуле указывает на то, что первый член суммы нужно разделить пополам.

Итак, метод интерполирования совокупности равноотстоящих исходных данных степенным рядом может быть описан следующим образом. Применив линейную поправку, приводящую пару концевых данных к нулю, мы представляем исходные данные в виде ряда Фурье по синусам согласно методу, разобранному в IV, 12. Если данные

свободны от помех (математические данные), то мы оставляем это разложение в исходном виде. Если данные содержат налагающиеся на них помехи, то мы обрываем ряд в надлежаще выбранном месте. Мы имеем теперь ряд Фурье по синусам, содержащий m членов с коэффициентами $b_1, b_2, \dots, b_m$. Этот ряд мы превращаем в бесконечный ряд, расположенный по полиномам Чебышева:

$$g(x) = \sum_{k=0}^{\infty} c_k T_k(x), \quad (5-14.6)$$

коэффициенты которого вычисляются по формулам

$$\begin{aligned} c_{2k} &= 2(-1)^k \left[-J_{2k}\left(\frac{\pi}{2}\right) b_1 + J_{2k}\left(3\frac{\pi}{2}\right) b_3 - J_{2k}\left(5\frac{\pi}{2}\right) b_5 + \dots \right], \\ c_{2k+1} &= 2(-1)^k \left[-J_{2k+1}(\pi) b_2 + J_{2k+1}(2\pi) b_4 - J_{2k+1}(3\pi) b_6 + \dots \right]. \end{aligned} \quad (5-14.7)$$

Значения функций Бесселя четного порядка для дуг, нечетнократных $\frac{\pi}{2}$, и функций Бесселя нечетного порядка для дуг, кратных π , вычисляются заранее (ср. табл. XI). Нахождение коэффициентов разложения c_k сводится тогда к умножению коэффициентов b_k на числовую матрицу.

При отсутствии помех мы можем сделать еще один шаг. Вычисление коэффициентов Фурье b_k и последующее вычисление c_k можно объединить в одну операцию. Берем наши равноотстоящие данные, полученные после разделения четной и нечетной частей функции $g(x)$, и непосредственно умножаем их на заранее составленную числовую матрицу, получая, таким образом, в один прием четные и нечетные коэффициенты c_k без какого-либо предварительного разложения в ряд Фурье (см. табл. XII).

Теоретически разложение (6) по полиномам Чебышева представляет собой бесконечный ряд. Однако хорошая сходимость ряда дает возможность ограничиться небольшим числом членов. Мы вычисляем конечную сумму $\nu + 1$ членов

$$\bar{g}(x) = \sum_{k=0}^{\nu} c_k T_k(x) \quad (5-14.8)$$

в заданных точках и устанавливаем, насколько велика погрешность, вызванная усечением ряда. Мы останавливаемся на том числе членов $k = \nu$, при котором остатки оказываются достаточно малыми. При наличии помех естественным критерием ограничения числа членов ряда служит то соображение, что точность интерполяции не должна превышать точности данных. Поэтому мы в ряду (8) ограничимся той степенью, при которой максимальный остаток в любой из данных точек остается на уровне средней погрешности исходных значений.

Этот метод получения хорошо сходящегося разложения в степенной ряд для последовательности равноотстоящих данных имеет то преимущество, что не нужно заранее знать, какой степени полином наиболее подходит для наших целей. Исходные данные определяют наиболее соответствующий полином для заданной степени точности. Осложнения, связанные с явлением Рунге, при этом ликвидируются, и мы получаем лучшее приближение, сравнимое с тем, которое получается при теоретически более желательном, но практически гораздо труднее осуществимом расположении исходных значений в нулях первого отброшенного полинома Чебышева $T_{n+1}(x)$. В нашем случае, т. е. при равноотстоящих исходных точках, мы все же можем воспользоваться преимуществами разложения по полиномам Чебышева, которое дает практически равномерное приближение на всем интервале.

15. Сходимость полиномиальной интерполяции равноотстоящих данных

Предыдущий параграф содержал решение задачи интерполяции полиномами в случае равноотстоящих исходных данных. Это решение дает средство для преодоления двух основных трудностей: исключения помех и устранения неравномерности погрешностей, которая может привести к сильным искажающим колебаниям вблизи обоих концов интервала. Обе трудности были успешно преодолены с помощью промежуточного использования ряда Фурье по синусам, который дает возможность эффективно отделить помехи от основной функции и распределяет погрешности одного и того же порядка величины равномерно по всему интервалу. Конечное число членов результирующего ряда синусов разлагается далее в бесконечный ряд полиномов Чебышева, в котором опять ограничиваются надлежащим числом членов. Известно, что полученный таким образом интерполяционный ряд сходится к заданной функции $f(x)$, если эта функция конечна, однозначна, кусочно непрерывна и не имеет в заданном интервале бесконечно большого числа колебаний.

Однако, несмотря на эти результаты, исследование Рунге о сходящемся или расходящемся характере простой равноотстоящей интерполяции сохраняет свое фундаментальное значение. Так как некоторые функции дают в этом случае сходящиеся, а другие — расходящиеся разложения, то возникает вопрос, можно ли *заранее* определить, какой тип функции приведет к одному и какой тип — к другому виду разложения. Определить это действительно возможно, если известны аналитические свойства функции $f(x)$. Мы покажем, что, если можно рассматривать $u = f(z)$ как функцию комплексной переменной $z = x + iy$, то поведение $f(z)$ в некоторой области около оси x однозначно определяет характер сходимости процесса равноотстоящей интерполяции для этой функции без какого-либо подробного исследования остатка. Единственное, что для этого нужно знать, это —

содержит ли или нет некоторая явно заданная овальная область вокруг оси x особые точки функции. Если функция $f(z)$ остается аналитической во всей этой области, включая границу, то интерполяционный процесс с равноотстоящими данными для этой функции будет равномерно сходиться в интервале $[-1, +1]$. Если где-либо в указанной области имеется особая точка функции, то сходимость сохраняется лишь в некотором промежутке интервала $[-1, +1]$, тогда как вне этого промежутка интерполяционный ряд расходится по мере бесконечного увеличения n .

Это свойство сходимости очень похоже на свойство ряда Тейлора, который также сходится в зависимости от аналитического характера функции $f(z)$ вне оси x , хотя бы мы интересовались функцией исключительно только лишь в вещественном интервале. Если мы опишем окружность из центра разложения радиусом, равным расстоянию до ближайшей особой точки, то ряд Тейлора сходится внутри этого круга и расходится вне круга, между тем как в точках, лежащих на самой окружности, он может либо сходиться, либо расходиться.

В нашем случае отличие состоит в том, что критическая область оказывается значительно более сложной, чем круг, хотя математически может быть строго определена. Будем исходить из следующей основной функции:

$$\frac{F_m(z) - F_m(x)}{z - x}, \quad (5-15.1)$$

где $F_m(x)$ есть фундаментальный полином, т. е. полином, составленный из корневых множителей

$$F_m(x) = (x - x_1)(x - x_2) \dots (x - x_m). \quad (5-15.2)$$

Здесь $x_1, x_2, \dots, x_m$ — точки интерполирования.

Но числитель дроби (1) делится на знаменатель, следовательно, функция (1), рассматриваемая как функция от x , должна быть полиномом степени $m - 1$. Это же остается верным, если мы разделим ее на $F_m(z)$ и рассмотрим функцию

$$\Phi(x, z) = \frac{F_m(z) - F_m(x)}{F_m(z)(z - x)} = \frac{1}{z - x} - \frac{F_m(x)}{F_m(z)} \frac{1}{z - x}. \quad (5-15.3)$$

Следовательно, функция $\Phi(x, z)$, рассматриваемая как функция от x , допускает *точную* интерполяцию полиномом независимо от того, как расположены точки x_i . Ряд, осуществляющий разложение, содержит m членов. Кроме того, второй член правой части выражения (3) не влияет на интерполяцию, так как он исчезает во всех точках x_i . Этот член можно рассматривать как *остаточный* член интерполяции.

Мы видим, таким образом, что функция

$$v(x) = \frac{1}{z - x} \quad (5-15.4)$$

обладает тем свойством, что если интерполировать ее степенным рядом для произвольно заданных точек, то остаток интерполяции может быть представлен явно выраженной функцией.

Мы ограничимся случаем *равноотстоящих* заданных точек и положим

$$\left. \begin{aligned} x_k &= \frac{k}{n} \quad (k = 0, \pm 1, \dots, \pm n), \\ \varphi_{2k+1} &= kx \frac{(k^2x^2 - 1)(k^2x^2 - 4)\dots(k^2x^2 - k^2)}{(2k+1)!} \end{aligned} \right\} \quad (5-15.5)$$

Тогда $m = 2n + 1$, а $F_m(x)$ оказывается пропорциональной $(2n + 1)$ -й функции Стирлинга

$$\begin{aligned} F_{2n+1}(x) &= x \left(x^2 - \frac{1}{n^2} \right) \left(x^2 - \frac{4}{n^2} \right) \dots \left(x^2 - \frac{n^2}{n^2} \right) = \\ &= \frac{(2n+1)!}{n^{2n+1}} \varphi_{2n+1}(x). \end{aligned} \quad (5-15.6)$$

Разложение функции (4) в степенный ряд принимает вид

$$\begin{aligned} \frac{1}{z-x} &= \frac{n}{\varphi_1(z)} + \frac{n\varphi_1(x)}{2\varphi_2(z)} + \dots + \frac{n}{2n+1} \frac{\varphi_{2n}(x)}{\varphi_{2n+1}(z)} + \eta_{2n+1}(x, z) = \\ &= n \sum_{k=0}^{2n} \frac{\varphi_k(x)}{(k+1)\varphi_{k+1}(z)} + \eta_{2n+1}(x, z), \end{aligned} \quad (5-15.7)$$

где

$$\eta_{2n+1}(x, z) = \frac{\varphi_{2n+1}(x)}{\varphi_{2n+1}(z)} \frac{1}{z-x}. \quad (5-15.8)$$

Теперь мы воспользуемся основной интегральной теоремой Коши

$$f(x) = \frac{1}{2\pi i} \oint \frac{f(z) dz}{z-x}, \quad (5-15.9)$$

где интеграл берется по замкнутому контуру, который включает точку $z=x$, а также заданные точки $z=x_i$, но не содержит каких-либо особых точек функции $f(z)$. Если в этой формуле заменить функцию (4) ее разложением (7), то мы получим в правой части разложение Стирлинга для $f(x)$ с остаточным членом. Этот остаток имеет следующий вид:

$$\eta_{2n+1}(x) = \frac{\varphi_{2n+1}(x)}{2\pi i} \oint \frac{f(z) dz}{(z-x)\varphi_{2n+1}(z)}. \quad (5-15.10)$$

Но функция $\varphi_{2n+1}(x)$ остается в пределах ± 1 на всем интервале $(-1, +1)$. Поэтому мы сосредоточим свое внимание на функции

$$\begin{aligned} \varphi_{2n+1}(z) &= \frac{(nz+n)(nz+n-1)\dots(nz-n)}{(2n+1)!} = \\ &= \frac{(nz+n)!}{(nz-n-1)!(2n+1)!} \end{aligned} \quad (5-15.11)$$

и исследуем ее поведение в случае, когда n бесконечно возрастает. Воспользуемся прежде всего формулой для факториальной функции¹⁾

$$\frac{1}{(-\nu-1)!} = -\nu! \frac{\sin \pi \nu}{\pi}, \quad (5-15.12)$$

на основании которой рассматриваемая функция принимает вид

$$\varphi_{2n+1}(z) = \frac{(-1)^n \sin n\pi z}{\pi} \frac{[n(1+z)]! [n(1-z)]!}{(2n+1)!}. \quad (5-15.13)$$

Условимся, что комплексная переменная z остается в положительной полуплоскости. Тогда мы можем эффективно аппроксимировать факториальную функцию, используя формулу Стирлинга

$$n! = \sqrt{2\pi n} n^{n+\frac{1}{2}} e^{-n}. \quad (5-15.14)$$

Эта формула становится сколь угодно точной при возрастании n до бесконечности, но даже в области малых значений n она замечательно точна.

При таком приближении наша функция принимает вид

$$\varphi_{2n+1}(z) = \frac{(-1)^n}{\sqrt{\pi}} \frac{\sqrt{n}}{2n+1} \sqrt{1-z^2} \left[\frac{(1+z)^{1+z} (1-z)^{1-z}}{4} \right]^n \sin n\pi z. \quad (5-15.15)$$

Последний множитель этого выражения требует более внимательного рассмотрения:

$$\sin [n\pi (x + iy)] = \frac{1}{2i} (e^{-n\pi y} e^{in\pi x} - e^{n\pi y} e^{-in\pi x}). \quad (5-15.16)$$

Предположим, что y — положительное число. Тогда первый член становится пренебрежимо малым при безграничном возрастании n , между тем как второй член мы можем объединить с предыдущим множителем в одно выражение, возведенное в n -ю степень.

Для наших целей достаточно исследовать *абсолютное значение* этого выражения:

$$A(z) = \frac{1}{4} |(1+z)^{1+z} (1-z)^{1-z}| e^{\pi y}. \quad (5-15.17)$$

Когда n бесконечно возрастает, n -я степень этого числа может изменяться по-разному. Если $A(z)$ меньше 1, то n -я степень стремится к нулю, если $A(z)$ больше 1, — она стремится к бесконечности. Соответственно этому подынтегральное выражение остаточного члена (10) в первом случае стремится к бесконечности, а во втором случае — к нулю. Граница $A(z) = 1$ изображается замкнутой эллипсо-

¹⁾ Ср. {15}, стр. 239.

образной кривой (см. рис. 35), определяемой трансцендентным уравнением

$$\frac{1}{2}(1+x)\log[(1+x)^2+y^2]+\frac{1}{2}(1-x)\log[(1-x)^2+y^2]+ \\ +\pi|y|-y\left(\operatorname{arctg}\frac{y}{1+x}+\operatorname{arctg}\frac{y}{1-x}\right)=2\log 2.$$

Нижеследующая таблица дает значения функции $y(x)$ через интервалы 0,1 независимой переменной. В окрестности особой точки $x=1$, $y=0$ вычисления сделаны через интервалы 0,01.

x	y	x	y	x	y
0,0	0,5255	0,6	0,3855	0,95	0,0963
0,1	0,5219	0,7	0,3283	0,96	0,0813
0,2	0,5110	0,8	0,2556	0,97	0,0652
0,3	0,4925	0,9	0,1598	0,98	0,0475
0,4	0,4660	1,0	0,0	0,99	0,0273
0,5	0,4307			1,00	0,0

Для любого замкнутого пути, который полностью лежит вне этой трансцендентной кривой C , интеграл правой части формулы (10) стремится к нулю при бесконечном возрастании числа n . Это значит, что интерполяция сходится к истинному значению $f(x)$ во всем интервале интерполирования. Однако такой замкнутый путь интегрирования в формуле (9) может быть выбран лишь в том случае, когда внутренность кривой C не содержит никакой особой точки аналитической функции $f(z)$. Если же $f(z)$ имеет хотя бы одну особую точку внутри этой замкнутой области, то остаток стремится к бесконечности, а не к нулю и, следовательно, интерполяция не может всюду сходиться.

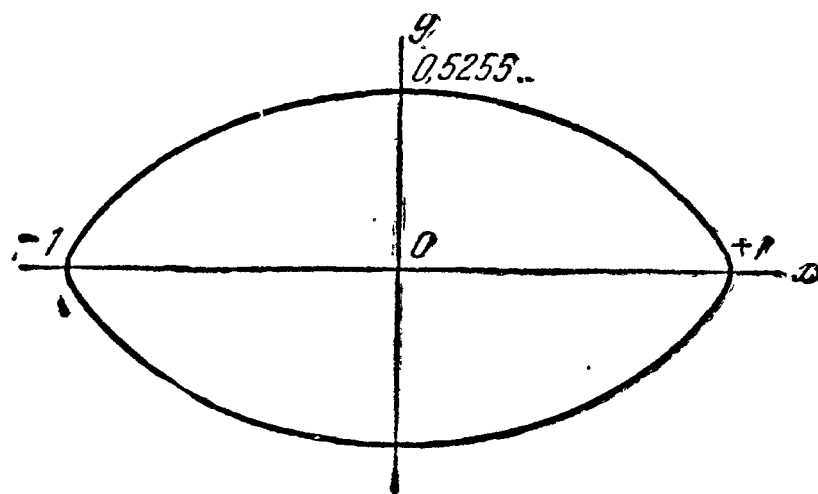


Рис. 35.

Таким образом, мы нашли необходимые и достаточные условия сходимости или расходимости равноотстоящей интерполяции. Рунге иллюстрировал расходимость равноотстоящей интерполяции полиномами на примере

$$y = \frac{1}{1+25x^2}.$$

Здесь особыми точками являются

$$z = \pm \frac{1}{5} i,$$

которые, очевидно, лежат внутри критической области.

16. Системы ортогональных функций *)

Гораздо более глубокое понимание природы интерполяционных задач было получено в связи с другой областью математики, развитой в течение XIX века, но полностью понятой во всех своих следствиях только в более позднее время. Остроумная *геометрическая* интерпретация метода наименьших квадратов открыла новое, исключительно благодарное поле для исследований. В этой геометрической картине понятие «функции» переходит в понятие «вектора», принадлежащего пространству *бесконечно большого* числа измерений.

Допустим, что задана непрерывная область переменного x , ограниченная значениями $x=a$, $x=b$ (области более чем одного измерения могут трактоваться совершенно аналогично). В этой области мы рассматриваем функцию $f(x)$, которая считается всюду конечной, однозначной и по крайней мере кусочно непрерывной. Такая функция, хотя и имеет бесконечное множество значений, может быть тем не менее табулирована в заданном интервале с любой степенью точности. Вместо того чтобы задать все значения функции $f(x)$, мы рассмотрим ее значения на всюду плотном множестве дискретных точек $x_1, x_2, \dots, x_n, \dots$. Кусочно непрерывный характер функции $f(x)$ дает возможность сделать так, чтобы дискретный ряд значений

$$y_1 = f(x_1), y_2 = f(x_2), \dots, y_n = f(x_n) \quad (5-16.1)$$

подходил сколь угодно близко к любому значению $f(x)$, так что это дискретное множество может заменить первоначальную функцию при всех аналитических операциях. Хотя при таком замещении и допускается известная погрешность, но она может быть сделана сколь угодно малой, если брать n достаточно большим.

Будем теперь рассматривать эти значения функции $f(x)$ как прямоугольные координаты точки в воображаемом евклидовом пространстве n измерений. Более точно, отложим на последовательных координатных осях значения

$$f_1 = y_1 \sqrt{\varepsilon_1}, f_2 = y_2 \sqrt{\varepsilon_2}, \dots, f_n = y_n \sqrt{\varepsilon_n}, \quad (5-16.2)$$

где

$$\varepsilon_k = x_k - x_{k-1} \quad (x_0 = a).$$

Все ε_k стремятся к нулю при бесконечном возрастании n .

*) Более подробное изложение вопросов, рассмотренных в этом и следующих параграфах гл. V, читатель найдет в книге [6в].

Мы, таким образом, получаем определенную точку

$$P = (f_1, f_2, \dots, f_n)$$

n -мерного пространства, изображающую заданную функцию $f(x)$.

Можно сказать также, что мы построили вектор $\overrightarrow{OP}$, проекции которого на координатные оси пропорциональны заданным значениям функции $f(x)$. Переходя ко все большему и большему n , мы получим все более адекватное представление любой конечной однозначной и кусочно непрерывной функции в заданном интервале. Аналитическое понятие функции может быть, следовательно, заменено более наглядным понятием *вектора* многомерного пространства. Квадрат длины этого вектора определяется суммой

$$f^2 = \sum_{i=1}^n f_i^2 = \sum_{i=1}^n y_i^2 \varepsilon_i. \quad (5-16.3)$$

Существенно заметить, что в пределе, когда n бесконечно возрастает, эта сумма заменяется интегралом

$$f^2 = \int_a^b f^2(x) dx. \quad (5-16.4)$$

Если рассматриваются две функции $f(x)$ и $g(x)$, то они представляют два вектора, взаимное расположение которых характеризуется их «скалярным произведением»

$$fg = \sum_{i=1}^n f_i g_i = \sum_{i=1}^n f(x_i) g(x_i) \varepsilon_i. \quad (5-16.5)$$

Оно также переходит в интеграл при бесконечном возрастании n :

$$fg = \int_a^b f(x) g(x) dx. \quad (5-16.6)$$

При такой интерпретации задача аппроксимирования функции с помощью заданной совокупности функций — например, с помощью степенных функций в задачах полиномиального интерполирования — также представляется в новом свете. Заданная совокупность функций представляет совокупность векторов, которую можно понимать как заданную *систему отсчета* в нашем воображаемом пространстве. Если число аппроксимирующих функций конечно, то это на самом деле только *частичная* система, так как число измерений бесконечно возрастает. Задачу аппроксимирования функции в виде линейной комбинации заданных функций $u_i(x)$ следует понимать теперь как геометрическую задачу *разложения вектора по заданным координатным осям*. Но тогда мы, очевидно, предпочтем, как наиболее удобную,

ортogonalную систему координат, которая характеризуется тем, что любые два базисных вектора u_i ортогональны друг другу:

$$u_i u_k = \int_a^b u_i(x) u_k(x) dx = 0, \quad (5-16.7)$$

а длина каждого из этих векторов нормирована к единице:

$$u_i^2 = \int_a^b u_i^2(x) dx = 1. \quad (5-16.8)$$

Функции, удовлетворяющие этим условиям, называются «ортонормальными». Факт нормирования к единице имеет меньшее значение, чем условие (7), которое уже само характеризует ортогональность заданной совокупности функций. В такой ортогональной системе отсчета задача разложения данного вектора разрешается непосредственно. Простое *проектирование* вектора v на ортогональные оси дает в этом частном случае координаты вектора v :

$$v = c_1 u_1 + c_2 u_2 + \dots + c_m u_m,$$

где

$$c_i = \frac{u_i v}{(u_i)^2}.$$

Это означает, что функция $f(x)$, разложенная на составляющие в системе отсчета, образованной функциями $u_i(x)$, представится в виде

$$f(x) = \sum_{i=1}^m c_i u_i(x), \quad (5-16.9)$$

где

$$c_i = \frac{\int_a^b f(x) u_i(x) dx}{\int_a^b u_i^2(x) dx}. \quad (5-16.10)$$

Это, однако, требует, чтобы $f(x)$ лежала в пространстве, порождаемом функциями $u_1(x), \dots, u_m(x)$. Если $f(x)$ лежит частью вне этого пространства, то мы получаем при этом построении функцию, которую можно считать эффективным *приближением* для $f(x)$, выраженным через функции $u_i(x)$. Такое приближение есть *проекция* функции $f(x)$ на подпространство базисных векторов $u_1(x), \dots, u_m(x)$. Эту проекцию можно рассматривать как такую специальную линейную комбинацию заданных векторов $u_i(x)$, которая наиболее близко

подходит к данной функции $f(x)$, поскольку ошибка приближения

$$\eta_m(x) = f(x) - f_m(x) \quad (5-16.11)$$

имеет наименьшую возможную длину.

Можно представить себе, что у нас имеются системы функций, которые удовлетворяют условию ортогональности (7) и к которым можно добавлять без конца все новые и новые функции, обладающие этим свойством. Такое бесконечное множество функций может включать в себя все функциональное пространство. Это значит, что если $f(x)$ есть произвольная конечная, однозначная и кусочно непрерывная функция в данном интервале, — мы добавим еще условие, что $f(x)$ в этом интервале имеет лишь конечное число максимумов и минимумов, — то можно составить приближения все более возрастающего порядка; при этом, хотя никогда *в точности* не будет иметь место равенство

$$f(x) = \sum_{i=1}^n c_i u_i(x),$$

сколь велико ни было бы число n , мы все же можем получить *в пределе*:

$$f(x) = \lim_{n \rightarrow \infty} \sum_{i=1}^n c_i u_i(x). \quad (5-16.12)$$

Такие системы функций называются «полными ортогональными системами функций». Разложения вида (12) с коэффициентами, определенными по формуле (10), называются «ортогональными разложениями». Они играют исключительную роль в физических и математических задачах нашего времени. Они не ограничиваются случаем функций одного переменного, но существуют также в случае функций любого числа переменных. Равным образом, область разложения не обязательно конечна. С надлежащими оговорками даже функции, заданные в бесконечной области, находят себе место в функциональном пространстве.

Функции, составляющие ряд Фурье, $\sin kx$, $\cos kx$, рассматриваемые в интервале $[-\pi, +\pi]$, были первым примером полной ортогональной системы функций. Ко времени их открытия понятие функционального пространства и более широкие обобщения понятия ортогональности не были еще осознаны. Теперь мы знаем, что функции Фурье представляют собой только *одну* особенно интересную ортогональную систему отсчета в функциональном пространстве. Но имеется бесконечно много других таких систем (они получаются в результате «поворота» системы первоначальных осей, рассматриваемой как «твердое тело»), которые также обладают замечательными свойствами функции Фурье аппроксимировать самые причудливые функции. Все эти функциональные системы аналитически эквивалентны в том

же смысле, что и все ортогональные системы отсчета n -мерного евклидова пространства, которые, как известно, обладают одинаковыми метрическими свойствами ввиду однородности пространства в каждом направлении.

17. Самосопряженные дифференциальные операторы

В гл. II мы видели, что каждая симметрическая матрица с несливающимися собственными значениями определяет ортогональную систему, осями которой служат главные оси матрицы; в этом случае всегда имеется необходимое число этих осей. В функциональном пространстве соответственно обильный источник полных ортогональных систем функций мы получаем при помощи некоторого класса линейных дифференциальных операторов, называемых «самосопряженными».

Важное матричное тождество

$$yAx - x\tilde{A}y = 0 \quad (5-17.1)$$

имеет свой аналог в теории линейных дифференциальных операторов. Пусть D есть произвольный линейный дифференциальный оператор в обыкновенных или частных производных. Тогда существует однозначно определенный оператор $\tilde{D}$, получающийся с помощью чисто алгебраических и дифференциальных операций, со следующим свойством:

$$v Du - u \tilde{D}v \equiv \frac{\partial p_1}{\partial x_1} + \dots + \frac{\partial p_n}{\partial x_n}, \quad (5-17.2)$$

где $p_1, p_2, \dots, p_n$ — некоторые функции от $x_1, x_2, \dots, x_n$.

Интегрируя это тождество по произвольному объему в пространстве $x_1, x_2, \dots, x_n$ и преобразуя в правой части интеграл по объему в интеграл по поверхности по теореме Гаусса *), мы получаем так называемое «тождество Грина», которое и является аналогом соотношения (1):

$$\int (v Du - u \tilde{D}v) d\tau = \text{интегралу по поверхности}. \quad (5-17.3)$$

До сих пор функции u и v были совершенно произвольны, не считая условий дифференцируемости, требуемых операторами D и $\tilde{D}$. Теперь допустим, что функция $u(x)$ подчинена некоторым более или менее сильным «краевым условиям» на граничной поверхности. Тогда мы всегда можем указать надлежаще выбранные краевые условия

*) См., например, [6г].

для $v(x)$ — так называемые «сопряженные краевые условия», которые обращают в нуль правую часть равенства (3); тогда мы получим

$$\int (v Du - u \tilde{D}v) d\tau = 0. \quad (5-17.4)$$

Внутри области интегрирования функции u и v остаются произвольными.

Может случиться, далее, что $\tilde{D} = D$; в этом случае мы говорим о «самосопряженном дифференциальном операторе». Кроме того, мы можем подчинить функцию u таким краевым условиям, что краевые условия для v окажутся *тождественными* с условиями для u . Тогда мы имеем «самосопряженный дифференциальный оператор с самосопряженными краевыми условиями». Такой оператор, рассматриваемый как оператор в функциональном пространстве *), является аналогом симметрической матрицы. Его главные оси определяют полную ортогональную систему функций. Эти главные оси определяются дифференциальным уравнением

$$Du = \lambda u \quad (5-17.5)$$

вместе с заданными краевыми условиями.

Уравнение (5) имеет решения лишь при определенном выборе значений λ , называемых собственными значениями оператора D . Так как, однако, функциональное пространство имеет бесконечное число измерений, то собственных значений также может быть бесконечно много. Они всегда вещественны и всегда дискретны, если дифференциальный оператор D не имеет каких-либо особенностей и область интегрирования конечна. Собственные значения можно расположить в порядке возрастания их модуля, начиная с наименьшего по модулю значения $\lambda = \lambda_1$.

Ортогональность двух решений, соответствующих двум различным собственным значениям λ_i и λ_k , следует из формулы (4), если подставить вместо u и v , эти два решения u_i и u_k :

$$(\lambda_i - \lambda_k) \int u_i u_k d\tau = 0. \quad (5-17.6)$$

Так как первый множитель отличен от нуля (если λ_i и λ_k различны), то второй множитель должен равняться нулю, что и выражает ортогональность этих функций относительно области интегрирования. В случае кратных собственных значений соответствующие им собственные элементы не будут автоматически ортогональны друг другу (хотя они ортогональны ко всем другим собственным элементам u_i), но могут быть ортогонализированы подбором надлежащих линейных комбинаций (ср. II, 9).

*) То есть этот оператор, действуя на векторы бесконечномерного эвклидова пространства, создает векторы того же пространства.

18. Дифференциальный оператор Штурма—Лиувилля

Можно построить бесконечно много различных самосопряженных дифференциальных операторов вместе с соответствующими краевыми условиями. Однако наиболее важные ортогональные системы функций, встречающиеся в прикладном анализе, связаны с дифференциальными операторами *второго порядка*. Это самый низкий порядок для самосопряженных дифференциальных операторов, ибо дифференциальные операторы первого порядка (по крайней мере, с вещественными коэффициентами) не могут быть самосопряженными. Мы ограничимся случаем одного переменного и, таким образом, рассмотрим только обыкновенные дифференциальные операторы. Наиболее общий обыкновенный линейный самосопряженный дифференциальный оператор второго порядка, — впервые исследованный Штурмом и Лиувиллем и потому часто называемый их именем, — имеет следующий вид:

$$Dy = \frac{d}{dx}(py') + qy. \quad (5-18.1)$$

Функции $p(x)$ и $q(x)$ пока еще произвольны, хотя $p(x)$ должна быть дифференцируема, и мы обычно принимаем, что она не меняет знака в заданном интервале.

Тождество Грина (17.3), соответствующее этому дифференциальному оператору, принимает вид

$$\begin{aligned} \int_a^b \left\{ v \left[\frac{d}{dx}(pu') + qu \right] - u \left[\frac{d}{dx}(pv') + qv \right] \right\} dx = \\ = [p(vu' - uv')]_a^b. \end{aligned} \quad (5-18.2)$$

Краевое условие может быть взято любое, лишь бы оно, примененное к u и v , обращало в нуль значение интеграла на границе, например

$$u(a) = u(b) = 0,$$

или

$$u'(a) = u'(b) = 0.$$

В качестве простого примера примем $a = -\pi$, $b = \pi$, $p = -1$, $q = 0$. Краевые условия выбираем следующим образом:

$$u(-\pi) = u(\pi), \quad u'(-\pi) = u'(\pi).$$

Собственные значения и собственные функции определяются дифференциальным уравнением

$$-u'' = \lambda u,$$

решением которого служит

$$u = A \cos \sqrt{\lambda} x + B \sin \sqrt{\lambda} x.$$

Краевые условия дают для λ значения

$$\lambda_k = k^2 \quad (k = 0, 1, 2, \dots),$$

а ввиду произвольности коэффициентов A и B каждому собственному значению соответствуют *две* независимые функции. Мы можем отделить их, выбрав

$$u_k = \cos kx, \quad \bar{u}_k = \sin kx.$$

Это — ортогональные функции, входящие в ряд Фурье: они получены здесь в результате решения простой задачи Штурма—Лиувилля.

Распространим теперь наши рассуждения на задачу нахождения собственных значений, соответствующую наиболее общему линейному дифференциальному оператору второго порядка. Рассмотрим уравнение

$$A(x) u''(x) + B(x) u'(x) + [C(x) + \lambda] u(x) = 0. \quad (5-18.3)$$

Умножим его на функцию $\rho(x)$, подобранную так, чтобы вновь получить оператор типа (1). Для этого нужно, чтобы

$$\rho(x) B(x) = [\rho(x) A(x)]';$$

отсюда получаем условие

$$\frac{\rho'}{\rho} = \frac{B - A'}{A}. \quad (5-18.4)$$

Теперь мы можем положить

$$p(x) = \rho(x) A(x) \quad (5-18.5)$$

и записать уравнение (3) в самосопряженной форме:

$$\frac{d}{dx} [p(x) u'(x)] + \rho(x) [C(x) + \lambda] u(x) = 0. \quad (5-18.6)$$

Применяя тождество Грина, получаем снова для граничного члена следующее выражение:

$$| p(x) (vu' - uv') |_a^b \quad (5-18.7)$$

и опять принимаем, что оно равно нулю в силу краевых условий. Единственное различие по сравнению с предыдущим случаем состоит в том, что «ортогональность» двух решений, соответствующих собственным значениям λ_i и λ_k , теперь выражается формулой

$$\int_a^b \rho(x) u_i(x) u_k(x) dx = 0. \quad (5-18.8)$$

Условие «ортогональности» такого вида называется «взвешенной ортогональностью», так как функцию $\rho(x)$, которая должна оставаться положительной в интервале интегрирования, можно интерпретировать как весовой множитель. Система функций, удовлетворяющих такому условию «взвешенной ортогональности», не отличается по

существо от системы обыкновенных ортогональных функций; в этом можно убедиться, если ввести новую переменную ξ , определенную равенством

$$d\xi = \rho(x) dx. \quad (5-18.9)$$

Взвешенная ортогональность по переменной x переходит теперь в обыкновенную ортогональность по переменной ξ .

Взвешенная ортогональность особенно полезна, если область интегрирования становится бесконечной (в одном или в обоих направлениях). Надлежащий выбор функции $\rho(x)$ может предупредить расходимость интеграла (8) в связи с бесконечностью пределов интегрирования. Благодаря этому искусственному приему использование пространства функций может быть распространено на случай бесконечного интервала.

19. Гипергеометрический ряд

Одним из многих глубоких открытий Гаусса было введение некоторого бесконечного ряда, так называемого гипергеометрического ряда. Он определяет функцию от x , зависящую одновременно от трех постоянных α , β , γ , которые могут принимать произвольные вещественные или комплексные значения с единственным ограничением, что γ не должно быть целым отрицательным числом. Гипергеометрическая функция, обыкновенно обозначаемая через $F(\alpha, \beta, \gamma, x)$, определяется следующим бесконечным рядом:

$$F(\alpha, \beta, \gamma; x) = 1 + \frac{\alpha\beta}{\gamma \cdot 1} x + \frac{\alpha(\alpha+1)\beta(\beta+1)}{\gamma(\gamma+1) \cdot 1 \cdot 2} x^2 + \\ + \frac{\alpha(\alpha+1)(\alpha+2)\beta(\beta+1)(\beta+2)}{\gamma(\gamma+1)(\gamma+2) \cdot 1 \cdot 2 \cdot 3} x^3 + \dots \quad (5-19.1)$$

Он сходится для всех $|x| < 1$ и расходится при $|x| > 1$, за исключением того случая, когда ряд обрывается после конечного числа членов. Почти все специальные функции математической физики, за исключением гамма-функции, находятся в определенном отношении к гипергеометрическому ряду, который охватывает широкий класс функций.

Гипергеометрическая функция удовлетворяет следующему дифференциальному уравнению второго порядка — так называемому дифференциальному уравнению Гаусса:

$$x(1-x)u'' + [\gamma - (\alpha + \beta + 1)x]u' - \alpha\beta u = 0. \quad (5-19.2)$$

Это дифференциальное уравнение второго порядка принадлежит к тому же типу, что и уравнение (18.3); следует только положить

$$A(x) = x(1-x); \quad B(x) = \gamma - (\alpha + \beta + 1)x; \quad C(x) = 0; \quad \lambda = -\alpha\beta.$$

Для того чтобы привести его к самосопряженному виду (18.6), нам надо получить $\rho(x)$ из условия (18.4); это дает

$$\rho(x) = x^{\gamma-1} (1-x)^{\alpha+\beta-\gamma} \quad (5-19.3)$$

и, следовательно,

$$p(x) = x^{\gamma} (1-x)^{\alpha+\beta+1-\gamma}. \quad (5-19.4)$$

20. Полиномы Якоби

Задача нахождения собственных значений приводит к ортогональной системе функций только в случае, когда рассматривается самосопряженный дифференциальный оператор с самосопряженными краевыми условиями. В предыдущем параграфе мы нашли множитель, преобразующий дифференциальное уравнение Гаусса в самосопряженное.

Теперь мы исследуем вопрос о краевых условиях. Область интегрирования мы ограничим интервалом $[0, 1]$. Тогда граничный член, согласно (18.7), будет

$$|x^{\gamma} (1-x)^{\delta} (vu' - uv')|_0^1, \quad (5-20.1)$$

где мы положили

$$\alpha + \beta + 1 - \gamma = \delta. \quad (5-20.2)$$

Примем, что γ и δ — положительные константы:

$$\gamma > 0, \delta > 0. \quad (5-20.3)$$

В этом случае граничный член (1) сам по себе всегда равен нулю, и, казалось бы, не дает нам возможности получить каких-либо краевых условий для $u(x)$. Однако на самом деле точки $x=0$ и $x=1$ являются *особыми точками* нашего дифференциального уравнения, в которых решение, вообще говоря, обращается в бесконечность. Требование, чтобы $u(x)$ оставалось *конечным* в этих двух точках, является само по себе краевым условием на двух концах интервала. Конечность в точке $x=0$ уже исключает одно из решений гауссова дифференциального уравнения и сводит нашу задачу к гипергеометрическому ряду. Условие конечности в точке $x=1$ требует дополнительных ограничений.

Заметим, что при любых заданных γ и δ мы можем еще свободно распоряжаться α или β . Далее, гипергеометрический ряд (19.1) имеет то замечательное свойство, что если принять α равным целому отрицательному числу $-n$, то он автоматически обрывается на члене со степенью x^n . Тогда

$$\alpha = -n, \quad \beta = n + \gamma + \delta - 1, \quad (5-20.4)$$

и мы получаем решение нашей задачи собственных значений в виде полиномов

$$P_n^{(\gamma, \delta)}(x) = F(-n, n + \gamma + \delta - 1, \gamma; x), \quad (5-20.5)$$

называемых полиномами Якоби. Они обладают замечательным свойством ортогональности относительно множителя веса (19.3):

$$\rho(x) = x^{\gamma-1} (1-x)^{\delta-1} \quad (5-20.6)$$

и образуют полную ортогональную систему функций при любом выборе γ и δ , удовлетворяющих условиям (3). Для приложений представляют особый интерес некоторые специальные значения постоянных γ и δ .

Ультрасферические полиномы. Выбор

$$\gamma = \delta \quad (5-20.7)$$

приводит к весовому множителю, симметрическому относительно средней точки $x = \frac{1}{2}$ интервала. В этом случае часто бывает полезно перенести начало осей координат в эту среднюю точку путем преобразования

$$x = \frac{1-\xi}{2}. \quad (5-20.8)$$

Новая переменная ξ изменяется от -1 до $+1$. Полученные таким образом полиномы, называемые «ультрасферическими», будут попеременно четными и нечетными полиномами соответственно четности или нечетности числа n . Если обозначить новую переменную вновь через x , то получим следующую формулу для определения ультрасферических полиномов:

$$P_n^{(\gamma)}(x) = F\left(-n, n+2\gamma-1, \gamma; \frac{1-x}{2}\right). \quad (5-20.9)$$

Весовой множитель ортогональности теперь равен

$$\rho(x) = (1-x^2)^{\gamma-1}. \quad (5-20.10)$$

Особый интерес представляют следующие частные случаи:

а) Полиномы Лежандра: $\gamma = 1$. Здесь множитель веса оказывается равным единице и взвешенная ортогональность переходит в обыкновенную ортогональность:

$$P_n(x) = F\left(-n, n+1, 1; \frac{1-x}{2}\right). \quad (5-20.11)$$

Мы встретимся с этими важными полиномами позже при рассмотрении задач о квадратурах (ср. VI, 10 и 19).

б) Полиномы Чебышева: $\gamma = \frac{1}{2}$. Здесь множитель веса $(1-x^2)^{-1/2}$, и преобразование $x = \cos \theta$ устраняет взвешенность. Тогда полином

$$T_n(x) = F\left(-n, n, \frac{1}{2}; \frac{1-x}{2}\right) \quad (5-20.12)$$

преобразуется в $\cos n\theta$, и мы получаем косинус функции ряда Фурье. Полиномы (12) имеют самое широкое применение (см. гл. VII) и являются наиболее эффективными при аппроксимировании произвольных функций.

с) Полиномы Чебышева второго рода: $\gamma = \frac{3}{2}$.

$$U_n(x) = (n+1) F\left(-n, n+2, \frac{3}{2}; \frac{1-x}{2}\right) = \frac{\sin(n+1)\theta}{\sin\theta} \quad (\text{если } x = \cos\theta). \quad (5-20.13)$$

Они удобны для полиномиального представления функции, которая принимает большие значения в окрестности точек $x = \pm 1$ и остается малой в остальной части интервала (ср. III, 7, см. также IV, 28).

d) Случай $\gamma = \infty$. Этот случай интересен ввиду его связи с рядом Тейлора, который, таким образом, может рассматриваться как предел ортогонального разложения. Кроме того, если одновременно x заменить на ξ по формуле

$$\xi = \sqrt{\gamma} x,$$

то интервал переменной ξ преобразуется в интервал от $-\infty$ до $+\infty$, а множитель веса (10) переходит в $e^{-\xi^2}$. Результирующие полиномы называются эрмитовыми:

$$H_n(x) = \lim_{\gamma \rightarrow \infty} \sqrt{4\gamma^n} F\left[-n, n+2\gamma-1, \gamma; \frac{1}{2}\left(1 - \frac{x}{\sqrt{\gamma}}\right)\right]. \quad (5-20.14)$$

е) Полиномы Лагерра. Среди полиномов Якоби с несимметрическим весом ($\gamma \neq \delta$) случай $\delta \rightarrow \infty$ представляет особый интерес. Он соответствует случаю $\beta \rightarrow \infty$. Но тогда разложение (19.1) гипергеометрического ряда показывает, что мы можем ввести новую переменную $\xi = \beta x$ и получить в пределе, при бесконечном возрастании β , функцию

$$\Phi(\alpha, \gamma, \xi) = 1 + \frac{\alpha}{\gamma \cdot 1} \xi + \frac{\alpha(\alpha+1)}{\gamma(\gamma+1) \cdot 1 \cdot 2} \xi^2 + \dots, \quad (5-20.15)$$

ряд для которой сходится при всех (вещественных или комплексных) значениях переменной ξ . С точки зрения ортогональности интервалом переменной ξ является $[0, \infty]$, а множитель веса (6)

$$\rho(\xi) = \xi^{\gamma-1} e^{-\xi}.$$

Выбор $\gamma = 1$ дает полиномы Лагерра

$$L_n(x) = n! \Phi(-n, 1; x). \quad (5-20.16)$$

Они ортогональны с весом e^{-x} в бесконечном интервале $[0, \infty]$. Мы встретились с этим классом полиномов в задаче обращения преобразования Лапласа в связи с рассмотрением импульсной реакции электрической сети (ср. IV, 30).

21. Интерполирование ортогональными полиномами

В §§ 14 и 15 мы рассмотрели те опасности, с которыми сопряжено равноотстоящее полиномиальное интерполирование. Как мы видели, погрешность произвольной интерполяции с помощью степенного ряда зависит от отношения двух полиномов

$$Q_n(x, z) = \frac{F_n(x)}{F_n(z)}. \quad (5-21.1)$$

Здесь $F_n(x)$ — фундаментальный полином, образованный из двучленных множителей $x - x_i$, где x_i — узлы интерполяции. При этом x есть некоторая точка интервала $[-1, +1]$, а z — некоторая точка комплексной плоскости. Если отношение (1) стремится к нулю при бесконечном возрастании n , то сходимость интерполяционного ряда обеспечена для всех значений x . Но в случае равноотстоящей интерполяции для x допустимы только такие значения, которые лежат вне определенной овальной области, охватывающей интервал интерполяции. Это в свою очередь приводит к тому, что $f(x)$ должна быть аналитической функцией не только в заданном интервале, но ее аналитический характер должен сохраняться повсюду внутри указанной овальной области.

Совсем иначе ведут себя ортогональные полиномы. Мы можем разложить по полной ортогональной системе функций, согласно формуле (16.12), произвольную функцию $f(x)$, которая удовлетворяет в интервале интерполяции условиям, гораздо более слабым, чем требование аналитичности, и которая вне этого интервала может быть даже не определена. Однако коэффициенты c_i этого разложения требуют вычисления определенных интегралов (16.10), которое на практике лишь редко удается провести. Поэтому имеет большое практическое значение то, что мы можем получить эквивалентное разложение с измененными коэффициентами c'_i , которое также сходится к $f(x)$ с возрастанием n и точность которого даже при конечном n несущественно хуже, чем точность разложения, образованного с помощью коэффициентов c_i . Такое разложение мы получаем в явном виде на основе процесса *интерполяции*, не требующего интегрирования.

Мы можем связать этот процесс с «классическим» рядом

$$f(x) = \sum_{i=1}^{\infty} c_i \mu_i(x), \quad (5-21.2)$$

ограничиваясь n членами его и рассматривая остаток этого разложения:

$$\eta_n(x) = f(x) - f_n(x) = \sum_{i=n+1}^{\infty} c_i u_i(x). \quad (5-21.3)$$

Если мы примем, что этот ряд сходится быстро, то можно оценить остаток, взяв лишь *первый член*. В этом случае можно принять, что

$$\eta_n(x) = c_{n+1} u_{n+1}(x), \quad (5-21.4)$$

а это значит, что погрешность равна нулю для корней первой опущенной ортогональной функции. Но это в свою очередь значит, что мы получим коэффициенты конечного разложения

$$f_n(x) = \sum_{i=1}^n c'_i u_i(x), \quad (5-21.5)$$

подобрав их так, чтобы этот ряд принимал значения функции $f(x)$ в нулевых точках первой опущенной ортогональной функции $u_{n+1}(x)$ (если только n таких точек могут быть найдены внутри области ортогональности):

$$\sum_{i=1}^n c'_i u_i(\lambda_k) = f(\lambda_k); \quad u_{n+1}(\lambda_k) = 0. \quad (5-21.6)$$

Это дает систему n линейных уравнений для определения коэффициентов c'_i . Хотя эти коэффициенты не будут, вообще говоря, совпадать с классическими коэффициентами (16.10), полученными интегрированием, все же погрешность приближения не будет велика, между тем как процедура вычисления проста и приводит прямо к цели.

Сходимость интерполяции мы получили ценою того ограничения, что узловые значения функции $f(x)$ должны теперь задаваться для заранее определенных *неравноотстоящих* точек $x = \lambda_k$, являющихся нулями первого опущенного ортогонального полинома¹⁾.

¹⁾ То, что $p_n(x)$ должны иметь n корней в интервале ортогональности, легко доказывается. Действительно, предположив, что это не так, мы имели бы

$$p_n(x) = (x - \lambda_1)(x - \lambda_2) \dots (x - \lambda_m) q(x) \quad (m < n),$$

где $q(x)$ не меняет знака между a и b . Но тогда

$$\int_a^b (x - \lambda_1) \dots (x - \lambda_m) p_n(x) \rho(x) dx$$

не может равняться нулю, ибо подынтегральная функция не меняет знака на всем интервале интегрирования. В то же время этот интеграл должен равняться нулю, так как функция $p_n(x)$ ортогональна всякому полиному более низкой степени (ср. дальнейший текст). Это противоречие доказывает сформулированное утверждение.

Фундаментальный полином, входящий в формулу (1), должен быть заменен n -м ортогональным полиномом $p_n(x)$ [счет этих полиномов начинается с $n=0$ и, следовательно, $u_{(n+1)}(x)$ в действительности есть $p_n(x)$]. Если мы образуем отношение (1), то станет ясно, что имеет место значительное улучшение сходимости. Переменная z больше не связана с овальной областью около оси x . Мы можем выбрать для z *любое* комплексное значение $x + iy$, сколь угодно близкое к оси x , и все же $Q_n(x, z)$ стремится к нулю.

Это утверждение мы можем непосредственно доказать элементарными рассуждениями для частного случая полиномов Чебышева. В этом случае

$$p_n(x) = \cos n\theta,$$

если преобразовать x в переменную θ по формуле

$$x = \cos \theta.$$

Теперь для любого комплексного значения x , фактически для любого x вне интервала ± 1 , угол θ становится комплексным. Но мы можем положить

$$\begin{aligned} \theta &= \varphi + i\psi, \\ \cos n\theta &= \frac{1}{2} (e^{in\varphi - n\psi} + e^{-in\varphi + n\psi}), \end{aligned}$$

а это выражение стремится к бесконечности с возрастанием n при любом значении $\psi \neq 0$.

Мы видим, следовательно, что аналитический характер функции $f(x)$ вне данного интервала не является теперь существенным. В действительности интерполяция сходится к $f(x)$ при любом значении x из данного интервала, независимо от аналитического характера функции $f(x)$ даже в заданном интервале, поскольку $f(x)$ принадлежит к классу функций «ограниченной вариации».

Интерполяция функций с помощью ортогональных полиномов (не обязательно типа Якоби) обладает еще некоторыми другими свойствами, значительно облегчающими технику вычислений. Одним из замечательных свойств ортогональных полиномов является то, что они удовлетворяют *рекуррентному* соотношению, связывающему три последовательных полинома. Это рекуррентное соотношение имеет следующую общую форму:

$$p_{n+1}(x) = (c_{n+1}x - a_n)p_n(x) - b_np_{n-1}(x). \quad (5-21.7)$$

Существование такого соотношения есть прямое следствие того факта, что произвольный полином $p_n(x)$ ортогонален всем предшествующим полиномам и поэтому также всякой степени x^α , $\alpha < n$, так как такая степень есть просто линейная комбинация полиномов $p_k(x)$ степени, меньшей чем n . Но отсюда можно заключить, что $p_n(x)$ вообще ортогонален *любому* полиному степени, меньшей чем n .

Обозначим наивысший коэффициент полинома $p_n(x)$, т. е. коэффициент при x^n , через μ_n . Тогда разность

$$p_{n+1}(x) - \frac{\mu_{n+1}}{\mu_n} x p_n(x)$$

не содержит степени x^{n+1} и является полиномом степени n . Этот полином может быть представлен как линейная комбинация полиномов $p_0(x), p_1(x), \dots, p_n(x)$:

$$p_{n+1}(x) - \frac{\mu_{n+1}}{\mu_n} x p_n(x) = \gamma_0 p_0(x) + \gamma_1 p_1(x) + \dots + \gamma_n p_n(x). \quad (5-21.8)$$

Умножив теперь обе части этого равенства на $\rho(x) p_m(x)$ ($m < n - 1$), мы в силу ортогональности получаем

$$\gamma_m \int_a^b \rho(x) p_m^2(x) dx = - \frac{\mu_{n+1}}{\mu_n} \int_a^b \rho(x) p_n(x) x p_m(x) dx.$$

Но $x p_m(x)$ есть полином степени, меньшей чем n , которому $p_n(x)$ ортогонален. Следовательно, все γ_m ($m < n - 1$) должны выпасть из правой части равенства (8), и ненулевыми коэффициентами будут только γ_n и γ_{n-1} . Таким образом, доказано существование «трехчленного рекуррентного соотношения» вида (7), причем получено явное выражение для коэффициента c_{n+1} :

$$c_{n+1} = \frac{\mu_{n+1}}{\mu_n}. \quad (5-21.9)$$

Предположим теперь, что ортогональные полиномы нормированы по длине:

$$\int_a^b \rho(x) p_k^2(x) dx = 1. \quad (5-21.10)$$

Если записать $p_k^2(x)$ в форме

$$p_k(x) (\mu_k x^k + \dots)$$

и принять в расчет то, что поставленные здесь точки замещают полином степени, не превышающей $n - 1$, то получим

$$\mu_k \int_a^b \rho(x) p_k(x) x^k dx = 1. \quad (5-21.11)$$

Умножим теперь равенство (7) на $p_{n-1}(x) \rho(x)$ и проинтегрируем обе части, приняв во внимание, что $x p_{n-1}(x)$ дает $\mu_{n-1} x^n$ плюс полином не выше степени $(n - 1)$:

$$b_n = c_{n+1} \mu_{n-1} \int_a^b x^n p_n(x) \rho(x) dx = c_{n+1} \frac{\mu_{n-1}}{\mu_n}.$$

Мы получаем, таким образом, явное выражение также и для b_n :

$$b_n = \frac{\mu_{n-1} \mu_{n+1}}{\mu_n^2} \quad (5-21.12)$$

и можем теперь записать равенство (7) в несколько иной форме, умножив его на $\frac{\mu_n}{\mu_{n+1}}$:

$$\beta_{n+1} p_{n+1}(x) = (x - \alpha_n) p_n(x) - \beta_n p_{n-1}(x), \quad (5-21.13)$$

где мы положили

$$\alpha_k = a_k \frac{\mu_k}{\mu_{k+1}},$$

$$\beta_k = \frac{\mu_{k-1}}{\mu_k}.$$

Эти рекуррентные соотношения можно рассматривать как последовательность линейных уравнений

$$\left. \begin{aligned} (\alpha_0 - x) y_0 + \beta_1 y_1 &= 0, \\ \beta_1 y_0 + (\alpha_1 - x) y_1 + \beta_2 y_2 &= 0, \\ \beta_2 y_1 + (\alpha_2 - x) y_2 + \beta_3 y_3 &= 0, \\ &\dots \end{aligned} \right\} \quad (5-21.14)$$

Эта бесконечная последовательность, однако, обрывается, если мы рассмотрим те частные значения $x = \lambda_k$, для которых $p_n(x)$ равны нулю:

$$y_n = p_n(\lambda_k) = 0 \quad (k = 1, 2, \dots, n).$$

Мы тогда получаем однородную систему n линейных уравнений, определитель которой должен равняться нулю:

$$\begin{vmatrix} \alpha_0 - \lambda & \beta_1 & & & \\ \beta_1 & \alpha_1 - \lambda & \beta_2 & & \\ & \beta_2 & \alpha_2 - \lambda & \beta_3 & \\ & & \dots & \dots & \dots \\ & & & \beta_{n-1} & \alpha_{n-1} - \lambda \end{vmatrix} = 0. \quad (5-21.15)$$

Мы пришли, таким образом, к регулярной задаче собственных значений для симметрической матрицы n -го порядка. Собственные значения $\lambda = \lambda_i$ являются корнями уравнения $p_n(\lambda) = 0$. Собственные решения — главные оси симметрической матрицы, которые автоматически ортогональны друг другу. Эти собственные решения суть

$$u_i = p_0(\lambda_i), \quad p_1(\lambda_i), \dots, p_{n-1}(\lambda_i).$$

коэффициенты c'_i в форме

$$c'_i = \sum_{\alpha=1}^n \rho_{\alpha}^2 f(\lambda_{\alpha}) p_i(\lambda_{\alpha}). \quad (5-21.20)$$

Последние равенства можно рассматривать как естественный аналог определяющих равенств (16.10) для коэффициентов c_i . Большое преимущество этих новых коэффициентов состоит в том, что они могут быть получены численно с помощью простого процесса *суммирования* (причем для соответствующих множителей можно заранее составить таблицу) без всякого интегрирования.

И здесь полиномы Чебышева имеют большие преимущества. Для них весовые множители ρ_{α}^2 становятся все равными, и следовательно, взвешенная ортогональность оказывается обыкновенной ортогональностью. Кроме того, коэффициенты $p_k(\lambda_{\alpha})$ здесь легко находятся, ибо это просто тригонометрические функции $\cos k\theta$ для углов θ , которые легко найти в таблицах (ср. IV, 16).

Другими частными случаями, которые могут представить интерес, являются полиномы Лежандра и Лагерра. Хотя корни этих полиномов вместе с весами ρ_i уже вычислены¹⁾, разработанной таблицы матриц $p_k(\lambda_i)$ в настоящее время еще нет.

¹⁾ Веса ρ_i^2 находятся в замечательном соотношении с весами квадратуры Гаусса. Пусть интерполяция функции $f(x)$ с помощью ортогональных полиномов служит для получения парэктического значения определенного интеграла

$$A = \int_a^b \rho(x) f(x) dx$$

(«квадратуры Гаусса», ср. VI, 10). Тогда из ортогональности всех $p_i(x)$ к $p_0(x) = p_0 = \text{const}$ следует, что в процессе интегрирования все члены разложения (5), за исключением первого, исчезают, и мы получаем

$$\bar{A} = c'_0 p_0 \int_a^b \rho(x) dx.$$

Но тогда, в силу формулы (20),

$$\bar{A} = p_0^2 \int_a^b \rho(x) dx \sum_{\alpha=1}^n \rho_{\alpha}^2 f(\lambda_{\alpha}). \quad (5-21.21)$$

Множитель перед знаком суммирования равен 1 в соответствии с нормированием $p_0(x)$. С другой стороны, для весов w_i гауссовой квадратуры

$$\bar{A} = \sum_{\alpha=1}^n w_{\alpha} f(\lambda_{\alpha})$$

имеются таблицы. Сопоставление с формулой (21) дает

$$\rho_i^2 = w_i.$$

ЛИТЕРАТУРА К ГЛАВЕ V

- [1] См. {4}, главы IV и V.
- [2] См. {8}, главы VI и IX.
- [3] Fort T., Finite Differences and Difference Equations (Oxford University Press, New York, 1948).
- [4] Milne-Thompson, Calculus of Finite Differences (Macmillan, London, 1933).
- [4a] Шиголов Б. М., Математическая обработка наблюдений, Физматгиз, М., 1960.
- [4б] Гельфонд А. О., Теория конечных разностей, Физматгиз, М., 1959.
- [4в] Марков А., Исчисление конечных разностей, Матезис, Одесса, 1911.
- [5] Стефенсон И. Ф., Теория интерполяции, ОНТИ, 1936.
- [6] Whittaker E. T. and Robinson G., A Short Course in Interpolation (Blackie & Sohn, London, 1923).
- [6a] Kuntzmann I., Methodes numériques (Interpolation-Derivees), Dunod, Paris, 1959.
- [6б] Гончаров В. Л., Теория интерполирования и приближения функций, М.—Л., 1954.
- [6в] Шилов Е. М., Математический анализ, Физматгиз, 1960.
- [6г] Дубнов Я. С., Основы векторного исчисления, ч. II, ГТТИ, 1950.
- [6д] Березин И. С. и Жидков Н. П., Методы вычислений, ч. I, Физматгиз, 1959.

Статьи:

- [7] Schoenberg I. J., Some Analytical Aspects of the Problem of Smoothing (Courant Anniversary Volume, Interscience Publishers, New York, 1948).
 - [7a] Clenshow, Curve Fitting with a Digital Computer, Computer Journal, 1960, 1, 2, № 4, 170—173.
-

ГЛАВА VI

МЕТОДЫ КВАДРАТУР

1. Исторические замечания

Задача нахождения площади привлекала научную мысль с давних пор. «Задача Дидоны», обычно формулируемая как задача охватить максимальную площадь цепью заданной длины, восходит к доисторическим временам. Определение площади более или менее сложной конфигурации играло важную роль у земледельческих народов древним вавилонянам, как и древним египтянам, были известны различные формулы, выраженные, конечно, больше словесно, чем алгебраически, для вычисления площадей прямолинейных фигур. Греки, используя свои большие геометрические познания, исследовали проблему гораздо более подробно. Нахождение площади круга оказалось особенно привлекательной задачей; она привела к установлению строгих методов теории пределов, называвшейся в древнее время «методом исчерпывания». Фактическое интегрирование было проведено Архимедом, который пользовался методом вписанных и описанных многоугольников, получая, таким образом, нижнюю и верхнюю границы площади, неограниченно приближающиеся друг к другу. После открытия исчисления бесконечно малых площади многих фигур оказалось возможным вычислять, опираясь на тот факт, что интегрирование и дифференцирование являются обратными процессами. Кроме того, простой метод трапеций древних был усовершенствован путем использования полиномов второй и высшей степеней для интерполяции равноотстоящих данных. Исключительно полезная формула, основанная на применении параболы второго порядка, была введена английским математиком Симпсоном (1743); ее обычно называют формулой Симпсона. Позднее Гаусс (1814) нашел остроумный и очень полезный метод вычисления площадей, основанный на свойствах полиномов Лежандра.

Шведский математик Фредгольм ввел в 1900 г. интегральные уравнения, которые в последующие десятилетия приобрели фундаментальное значение для решения краевых задач, задач собственных значений и многих проблем современной статистики. Если

встречающиеся при этом интегралы вычислять по простой формуле трапеций, то интегральное уравнение переходит в систему обыкновенных линейных алгебраических уравнений с большим числом неизвестных. Но часто получаются более удобные решения, если к определенному интегралу, входящему в левую часть интегрального уравнения, применить более совершенные методы квадратуры.

2. Квадратура при помощи планиметров

Задача определения площади фигуры, ограниченной заданной кривой, часто называется «механической квадратурой», хотя при этом не всегда подразумевается, что при решении задачи используется какой-нибудь механический инструмент. В действительности, однако, существуют механические инструменты, которые выполняют операцию квадратуры с помощью механических или электрических средств. Эти инструменты называются «планиметрами», если их устройство основано на том принципе, что острие обводится по контуру измеряемой площади. Счетчик инструмента указывает непосредственно площадь в квадратных единицах.

Тщательно изготовленные планиметры являются довольно дорогими инструментами, и точность их ограничена. Они требуют тщательного вычерчивания и обведения контура измеряемой площади. Простое вычисление часто дает более точные результаты, чем планиметр, особенно в тех случаях, когда эмпирическая кривая задается не в виде непрерывной линии, а в виде ряда дискретных ординат. При вычислении используются только измеренные ординаты и не требуется, чтобы между заданными точками была произведена интерполяция каким-либо более или менее произвольным графическим приемом.

3. Правило трапеций

Старейшим методом аппроксимирования площади «под непрерывной кривой» является метод вписанных многоугольников, называемый в настоящее время «правилом трапеций». Мы соединяем точки, ординаты которых заданы, отрезками прямых и заменяем площадь под кривой площадью, расположенной под этой ломаной линией. Если заданные ординаты

$$y = y_0, y_1, y_2, \dots, y_n \quad (6-3.1)$$

соответствуют значениям абсциссы

$$x = x_0, x_1, x_2, \dots, x_n, \quad (6-3.2)$$

то элементарная формула площади трапеций дает следующее

выражение для площади многоугольника¹⁾:

$$\begin{aligned}\bar{A} &= \frac{1}{2} (y_0 + y_1) (x_1 - x_0) + \dots + \frac{1}{2} (y_{n-1} + y_n) (x_n - x_{n-1}) = \\ &= \frac{1}{2} [y_0 (x_1 - x_0) + y_1 (x_2 - x_0) + \dots + y_{n-1} (x_n - x_{n-2}) + \\ &\quad + y_n (x_n - x_{n-1})]. \quad (6-3.3)\end{aligned}$$

Для случая равноотстоящих абсцисс,

$$x_1 - x_0 = x_2 - x_1 = x_3 - x_2 = \dots = x_n - x_{n-1} = h, \quad (6-3.4)$$

формула (3) принимает более простой вид:

$$\bar{A} = h \left(\frac{1}{2} y_0 + y_1 + y_2 + \dots + y_{n-1} + \frac{1}{2} y_n \right). \quad (6-3.5)$$

При вычислении по этой формуле следует просто просуммировать все замеренные ординаты, приписав вес $\frac{1}{2}$ двум крайним ординатам.

Из основной теоремы интегрального исчисления известно, что площадь кривой, заданной уравнением $y = f(x)$, можно определить как предел, к которому стремится $\bar{A}$ при приближении h к нулю:

$$A = \int_a^b f(x) dx = \lim_{h \rightarrow 0} h \left(\frac{1}{2} y_0 + y_1 + \dots + y_{n-1} + \frac{1}{2} y_n \right). \quad (6-3.6)$$

Для теоретических целей этот предельный процесс вполне удовлетворителен, ибо для функций, заданных аналитически, мы часто в состоянии получить предел средствами анализа. Архимед при вычислении центра тяжести и метацентра многих сложных тел использовал именно предельный переход.

С точки зрения практического вычисления правило трапеций дает вполне удовлетворительные результаты, если мы располагаем достаточным числом ординат. Для применения этого правила требуется лишь непосредственное сложение заданного ряда ординат на арифмометре, которое представляет собой простой и быстрый процесс. Однако обеспечение большого набора ординат, требуемого при методе трапеций для достаточно хорошего приближения, является часто обременительным.

4. Правило Симпсона

Прямые линии слишком «жестки» для вполне удовлетворительного приближения кривых. Если мы хотим имитировать кривую, проводя ряд отрезков прямой, то следует для этой цели провести большое

¹⁾ Мы будем использовать далее черточку для обозначения «приближения». Следовательно, $\bar{A}$ есть приближение точной площади A ; аналогично $\bar{\eta}$ есть приближенное значение для η .

число малых отрезков. Мы, очевидно, поступим гораздо лучше, если для приближения воспользуемся *параболами второго порядка*. С помощью таких парабол можно аппроксимировать кривую очень хорошо, располагая аппроксимирующие параболы не слишком часто (ср. приближения отрезками *I* и дугами парабол *II* на рис. 36).

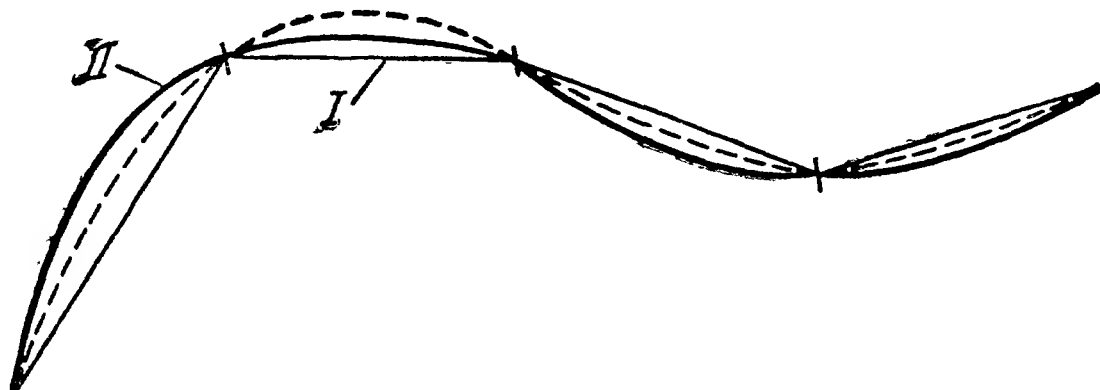


Рис. 36.

Большое число очень коротких прямых отрезков заменяется, таким образом, малым числом более длинных отрезков парабол.

Следовательно, мы значительно улучшим точность нашей квадратурной формулы, если вместо того, чтобы соединять две последова-

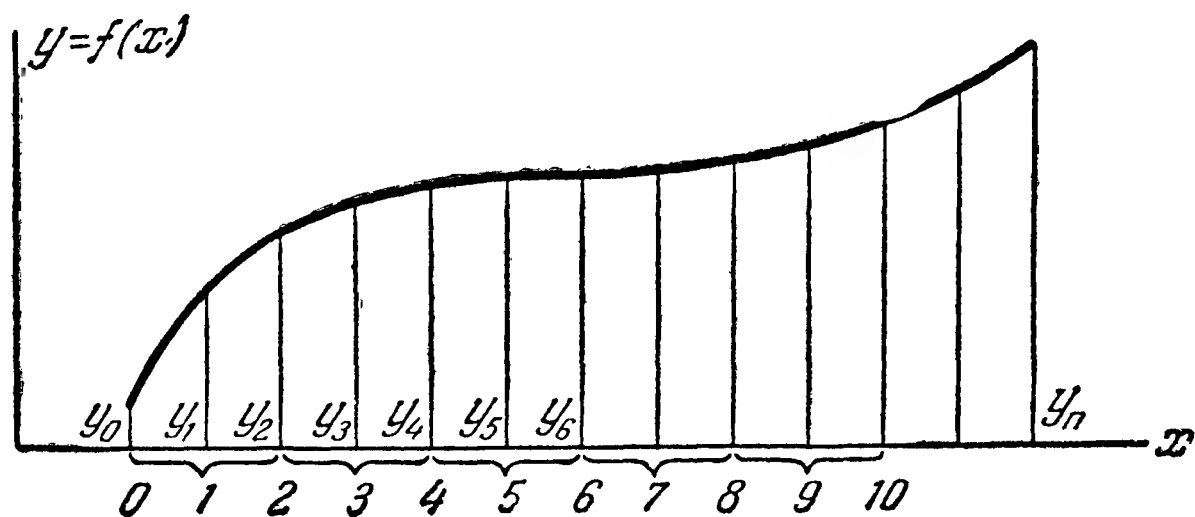


Рис. 37.

тельные точки прямыми, мы соединим три последовательные точки параболой. Всегда можно найти параболу вида

$$y = a + bx + cx^2, \quad (6-4.1)$$

которая проходит через концы трех заданных ординат. Таким путем мы подходим гораздо ближе к действительной кривой, чем с помощью простых отрезков прямой. Совершаемая ошибка поэтому будет гораздо меньшей.

Получающийся таким образом процесс известен как метод Симпсона, а окончательная формула называется формулой Симпсона. Разобьем всю площадь на четное число полос равной ширины, откладывая, следовательно, нечетное число ординат, так как обе конечные ординаты включаются в наш подсчет. Удобства ради примем ширину каждой полосы за единицу длины. В каждой двойной полосе аппроксимируем исходную кривую параболой второго порядка (рис. 37).

Рассмотрим, например, первую двойную полосу, содержащую три ординаты: y_0, y_1, y_2 . Мы можем разложить функцию $y=f(x)$ в окрестности точки $x=1$ в степенной ряд, пользуясь методом центральных разностей (см. V, 3). Согласно формуле Стирлинга имеем

$$f(1+t) = f(1) + \delta f(1)t + \frac{\delta^2 f(1)}{2} t^2, \quad (6-4.2)$$

где

$$\left. \begin{aligned} f(1) &= y_1, \\ \delta f(1) &= y_1 - y_0, \\ \delta^2 f(1) &= y_2 - 2y_1 + y_0. \end{aligned} \right\} \quad (6-4.3)$$

Таким образом, кривая в промежутке между $x=0$ и $x=2$ аппроксимирована параболой второго порядка, которая совпадает с заданной кривой в трех точках интерполяции: $x=0, 1, 2$.

Площадь полосы, которая ограничена интерполирующей параболой, можно получить, вычислив интеграл в пределах от $x=0$ до $x=2$:

$$\begin{aligned} \bar{A}_{02} &= \int_{-1}^{+1} f(1+t) dt = \left| f(1)t + \delta f(1) \frac{t^2}{2} + \frac{\delta^2 f(1)}{2} \frac{t^3}{3} \right|_{-1}^{+1} = \\ &= 2y_1 + \frac{1}{3} (y_2 - 2y_1 + y_0) = \frac{1}{3} y_0 + \frac{4}{3} y_1 + \frac{1}{3} y_2. \end{aligned} \quad (6-4.4)$$

Точно такой же процесс повторяем для площадей $\bar{A}_{24}, \bar{A}_{46}, \dots$ до тех пор, пока не исчерпаем всю искомую площадь. Мы получим при этом

$$\begin{aligned} \bar{A} &= \frac{1}{3} y_0 + \frac{4}{3} y_1 + \frac{1}{3} y_2 + \frac{1}{3} y_2 + \frac{4}{3} y_3 + \frac{1}{3} y_4 + \dots = \\ &= \frac{2}{3} \left(\frac{1}{2} y_0 + y_2 + y_4 + \dots + \frac{1}{2} y_{2n} \right) + \\ &\quad + \frac{4}{3} (y_1 + y_3 + y_5 + \dots + y_{2n-1}). \end{aligned} \quad (6-4.5)$$

Это и есть *формула Симпсона*. В ней отделены четные и нечетные ординаты и им приписаны различные веса, в отличие от метода трапеций, при котором все ординаты, кроме крайних, были взяты с одинаковым весом. Это различие в весах значительно увеличивает точность результата.

Если ширина каждой полосы составляет не 1, а h , надо просто умножить результат на h , и окончательная формула принимает вид

$$\bar{A} = \frac{2}{3} h \left[\frac{1}{2} y_0 + y_2 + \dots + \frac{1}{2} y_{2n} + 2(y_1 + y_3 + \dots + y_{2n-1}) \right]. \quad (6-4.6)$$

Необходимость четного числа полос иногда бывает неудобным ограничением при применении формулы Симпсона. Это затруднение

можно обойти, используя другое построение для *последних трех* (или *первых трех*) полос в том случае, когда число полос оказывается нечетным. Рассмотрим четыре ординаты y_0, y_1, y_2, y_3 . Поместим начало нашей системы отсчета в точку $x=1,5$ и аппроксимируем функцию

$$\frac{1}{2} [f(1,5+t) + f(1,5-t)]$$

параболой

$$y = a + ct^2,$$

учитывая, что площадь кривой между $t = \pm 1,5$ зависит лишь от четной части квадратной функции. По значениям функции при $t = 0,5$ и $1,5$ определяем коэффициенты a и c :

$$c = \frac{1}{4} [y_0 + y_3 - (y_1 + y_2)],$$

$$a = \frac{1}{16} [9(y_1 + y_2) - (y_0 + y_3)],$$

и для площади первых трех полос получаем выражение

$$\begin{aligned} \bar{A}_{03} &= \left| at + \frac{c}{3} t^3 \right|_{-1,5}^{+1,5} = \frac{3}{2} \cdot 2 \left(a + \frac{c}{3} \frac{9}{4} \right) = \\ &= \frac{3}{8} [3(y_1 + y_2) + (y_0 + y_3)]. \end{aligned} \quad (6-4.7)$$

Соответствующая формула для полос ширины h будет иметь вид

$$\bar{A}_{03} = \frac{3h}{8} [3(y_1 + y_2) + (y_0 + y_3)]. \quad (6-4.8)$$

Таким образом, если число полос нечетно, мы применяем формулу (8) к первым трем или последним трем полосам; тогда оставшееся число полос четно, и для них вступает в силу формула Симпсона (6).

5. Точность формулы Симпсона

Мы можем оценить точность параболического приближения на основании следующих соображений. При почленном интегрировании формулы Стирлинга в пределах ± 1 все члены с нечетными разностями выпадают, так как интеграл от нечетной функции, взятый между симметричными относительно начала пределами, равен нулю. Если принять теперь, что исходные ординаты расположены достаточно густо, то разложение Стирлинга будет сходиться настолько быстро, что можно оценить ошибку от отбрасывания дальнейших членов путем приравнивания ее первому опущенному члену. Первый член, которым мы пренебрегли при интегрировании (4.4), есть

$$\frac{\delta^4 f(1)}{24} \int_{-1}^{+1} t^2 (t^2 - 1) dt = -\frac{\delta^4 f(1)}{90}. \quad (6-5.1)$$

Эти же рассуждения справедливы и для всякой двойной полосы от $k=1$ до $k=2n-1$. Следовательно, для полной площади мы получаем

$$\int_0^{2n} f(x) dx = \bar{A} - \frac{1}{90} \sum_{k=0}^{n-1} \delta^4 f(2k+1). \quad (6-5.2)$$

Это выражение для погрешности не очень удобно, так как требует составления обширной таблицы разностей. Однако если $f(x)$ — достаточно гладкая функция, можно заменить разностные коэффициенты производными, а сумму — интегралом. Тогда оценка погрешности $\bar{\eta}$ выполненной квадратуры принимает следующий вид:

$$\bar{\eta} = -\frac{h^4}{180} [f'''(b) - f'''(a)], \quad (6-5.3)$$

где a есть нижний, b — верхний пределы интегрирования, а ширина полос принята равной h (о более общей оценке погрешности, не предполагающей гладкость функции, см. (17.6)).

Более удобен и надежен другой метод оценки погрешности, который будет рассмотрен позже (см. § 12). Если применить этот метод к квадратурной формуле Симпсона, то получим следующую оценку погрешности:

$$\bar{\eta} = \frac{4h}{15} \left[\left(\frac{1}{2} y_0 + y_2 + \dots + \frac{1}{2} y_{2n} \right) - (y_1 + y_3 + \dots + y_{2n-1}) \right] - \frac{h^2}{15} [f'(b) - f'(a)]. \quad (6-5.4)$$

6. Точность правила трапеций

Для того чтобы оценить точность правила трапеций, мы воспользуемся формулой Бесселя, которая интерполирует на половинных строках:

$$f\left(\frac{1}{2} + t\right) = \delta^0 f\left(\frac{1}{2}\right) + \delta^1 f\left(\frac{1}{2}\right) t + \frac{\delta^2 f\left(\frac{1}{2}\right)}{2} \left(t^2 - \frac{1}{4}\right) + \dots \quad (6-6.1)$$

Интегрируя между пределами $t = \pm \frac{1}{2}$, мы получаем площадь одной полосы

$$\begin{aligned} A_{01} &= \int_{-\frac{1}{2}}^{+\frac{1}{2}} f\left(\frac{1}{2} + t\right) dt = \\ &= \left| \delta^0 f\left(\frac{1}{2}\right) t + \delta^1 f\left(\frac{1}{2}\right) \frac{t^2}{2} + \frac{\delta^2 f\left(\frac{1}{2}\right)}{2} \left(\frac{t^3}{3} - \frac{t}{4}\right) \right|_{-\frac{1}{2}}^{+\frac{1}{2}}. \end{aligned} \quad (6-6.2)$$

Суммируя площади всех полос и пренебрегая членами высших степеней, получаем

$$A = \sum_{k=0}^{n-1} \delta^0 f\left(k + \frac{1}{2}\right) - \frac{1}{12} \sum_{k=0}^{n-1} \delta^2 f\left(k + \frac{1}{2}\right) = \\ = \bar{A} - \frac{1}{12} [\delta f(n) - \delta f(0)]. \quad (6-6.3)$$

Для случая полос ширины h формула принимает вид

$$A = \bar{A} - \frac{h}{12} [\delta f(nh) - \delta f(0)], \quad (6-6.4)$$

а заменив снова разности производными, мы получаем окончательную формулу, соответствующую формуле (5.3):

$$\bar{\eta} = -\frac{h^2}{12} [y'(b) - y'(a)]. \quad (6-6.5)$$

Сопоставление формул (5.3) и (6.5) показывает, что точность формулы Симпсона, вообще говоря, превышает точность формулы трапеций. Вторая степень h заменяется четвертой степенью, в то время как численный множитель 12 в знаменателе заменяется гораздо большим множителем 180. С другой стороны, однако, первая производная заменяется третьей производной, которая часто бывает менее гладкой, чем первая.

7. Правило трапеций с концевой поправкой

Мы можем уяснить себе содержание правила трапеций еще с другой точки зрения, если введем в рассмотрение ряд Фурье. Допустим, что функция $f(x)$ задана в интервале $(0, 1)$; разложим ее в ряд Фурье по косинусам:

$$f(x) = \frac{1}{2} a_0 + a_1 \cos \pi x + a_2 \cos 2\pi x + \dots \quad (6-7.1)$$

Подставим вместо x значения

$$0, \frac{1}{n}, \frac{2}{n}, \dots, \frac{n}{n}$$

и выполним суммирование:

$$\bar{A} = \frac{1}{n} \left[\frac{1}{2} f(0) + f\left(\frac{1}{n}\right) + \dots + f\left(\frac{n-1}{n}\right) + \frac{1}{2} f\left(\frac{n}{n}\right) \right]. \quad (6-7.2)$$

Тогда получим равенство

$$\bar{A} = \frac{1}{2} a_0 + a_{2n} + a_{4n} + \dots \quad (6-7.3)$$

Но по общей формуле, определяющей коэффициенты Фурье:

$$a_k = 2 \int_{-1}^1 f(x) \cos k\pi x dx. \quad (6-7.4)$$

Следовательно, $\frac{1}{2} a_0$ есть истинная площадь A кривой, и мы получаем

$$A = \bar{A} - (a_{2n} + a_{4n} + \dots). \quad (6-7.5)$$

Мы видим, таким образом, что погрешность правила трапеций будет тем меньшей, чем лучше сходится ряд Фурье (1). Но, как известно, сходимость ряда Фурье определяется аналитическим поведением функции. Чем более непрерывна сама функция и ее производные, тем лучше сходится ряд Фурье. В рассматриваемом нами случае ряда косинусов мы имеем во всем интервале $[-1, 1]$ четную функцию, которая остается непрерывной на концах, между тем как производная будет, вообще говоря, претерпевать разрыв в точках $x=0$ и $x=1$. Порядок величин коэффициентов a_k будет определяться этим разрывом. Если в (4) мы проинтегрируем по частям, то получим

$$\begin{aligned} a_k &= 2 \left| \frac{f(x) \sin k\pi x}{k\pi} \right|_0^1 - \frac{2}{k\pi} \int_0^1 f'(x) \sin k\pi x dx = \\ &= \frac{2}{(k\pi)^2} \left| f'(x) \cos k\pi x \right|_0^1 - \frac{2}{(k\pi)^2} \int_0^1 f''(x) \cos k\pi x dx. \end{aligned} \quad (6-7.6)$$

Следовательно, коэффициенты с четным индексом будут равны

$$a_{2k} = \frac{1}{2k^2\pi^2} [f'(1) - f'(0)] + \varepsilon, \quad (6-7.7)$$

где ε по сравнению с первым членом становится малым, когда k делается большим. Если, однако, случится, что выполняется граничное условие

$$f'(0) = f'(1), \quad (6-7.8)$$

то мы можем еще раз повторить интегрирование по частям и получить

$$a_{2k} = -\frac{1}{8k^4\pi^4} [f'''(1) - f'''(0)] + \varepsilon. \quad (6-7.9)$$

В первом случае равенство (5) дает

$$A = \bar{A} - [f'(1) - f'(0)] \frac{1}{2n^2\pi^2} \left[1 + \frac{1}{2^2} + \frac{1}{3^2} + \dots \right] + \varepsilon, \quad (6-7.10)$$

а во втором

$$A = \bar{A} + \frac{f'''(1) - f'''(0)}{8n^4\pi^4} \left[1 + \frac{1}{2^4} + \frac{1}{3^4} + \dots \right] + \varepsilon. \quad (6-7.11)$$

Бесконечные суммы, появляющиеся в этих равенствах, могут быть выражены через числа Бернулли B_{2k} , если воспользоваться соотношением

$$B_{2k} = \frac{2(2k)!}{(2\pi)^{2k}} \left(1 + \frac{1}{2^{2k}} + \frac{1}{3^{2k}} + \dots \right). \quad (6-7.12)$$

Первые числа Бернулли равны

$$B_2 = \frac{1}{6}, \quad B_4 = \frac{1}{30}, \quad B_6 = \frac{1}{42}, \quad \dots$$

Следовательно,

$$\begin{aligned} \frac{1}{\pi^2} \left(1 + \frac{1}{2^2} + \frac{1}{3^2} + \dots \right) &= \frac{1}{6}, \\ \frac{1}{\pi^4} \left(1 + \frac{1}{2^4} + \frac{1}{3^4} + \dots \right) &= \frac{1}{90}. \end{aligned} \quad (6-7.13)$$

Таким образом, из формулы (10) мы получаем следующее выражение для оценки погрешности правила трапеций:

$$A = \bar{A} - \frac{1}{12n^2} [f'(1) - f'(0)] = \bar{A} - \frac{h^2}{12} [f'(1) - f'(0)] \quad (6-7.14)$$

в согласии с нашим прежним результатом (6.5).

Однако надлежащим видоизменением функции $f(x)$ мы можем устранить скачок в первой производной и взять в основу расчетов формулу (11). Рассмотрим для этого функцию

$$g(x) = f(x) - \frac{f'(1) - f'(0)}{2} x^2. \quad (6-7.15)$$

Для этой функции выполняется граничное условие (8) (если заменить f на g), и мы получаем оценку [ср. (11) и (13)]

$$\int_0^1 g(x) dx = \bar{A}_g + \frac{g'''(1) - g'''(0)}{720n^4} = \bar{A}_g + \frac{h^4}{720} [f'''(1) - f'''(0)]. \quad (6-7.16)$$

Но, по определению функции $g(x)$,

$$\int_0^1 g(x) dx = \int_0^1 f(x) dx - \frac{1}{6} [f'(1) - f'(0)], \quad (6-7.17)$$

а операция $\bar{A}_g$ состоит из следующих двух частей:

$$\begin{aligned} \bar{A}_g = \frac{1}{n} \left[\frac{1}{2} f(0) + f\left(\frac{1}{n}\right) + \dots + f\left(\frac{n-1}{n}\right) + \frac{1}{2} f\left(\frac{n}{n}\right) \right] - \\ - \frac{f'(1) - f'(0)}{2n^3} \sum_{k=0}^n k^2, \end{aligned} \quad (6-7.18)$$

$$\sum_{k=0}^n k^2 = \frac{n(n+1)(2n+1)}{6} - \frac{n^2}{2} = \frac{n^3}{3} + \frac{n}{6}. \quad (6-7.19)$$

Таким образом, мы получаем

$$\int_0^1 f(x) dx = \frac{1}{n} \left[\frac{1}{2} f(0) + f\left(\frac{1}{n}\right) + \dots + \frac{1}{2} f(1) \right] - \frac{1}{12n^2} [f'(1) - f'(0)] + \frac{1}{720n^4} [f'''(1) - f'''(0)]. \quad (6-7.20)$$

Возвращаясь к нашим первоначальным обозначениям, имеем

$$\bar{A} = h \left[\frac{1}{2} y_0 + y_1 + \dots + y_{n-1} + \frac{1}{2} y_n \right] - \frac{h^2}{12} [f'(b) - f'(a)] \quad (6-7.21)$$

с оценкой погрешности

$$\bar{\eta} = \frac{h^4}{720} [f'''(b) - f'''(a)]. \quad (6-7.22)$$

Этот результат показывает, что поправка

$$- \frac{h^2}{12} [f'(b) - f'(a)] \quad (6-7.23)$$

(которая требует знания производной в двух конечных точках интервала) намного увеличивает точность простого правила трапеций. Новая погрешность, если сравнить ее с оценкой погрешности (5.3) формулы Симпсона, составляет лишь $1/4$ этого значения и имеет противоположный знак.

Применение правила трапеций с концевой поправкой будет особенно полезным, если заданные ординаты являются результатом наблюдений и, следовательно, подвержены случайным ошибкам. Образование простого арифметического среднего после предварительного уменьшения вдвое двух крайних ординат будет иметь тенденцию свести к минимуму эти ошибки. Кроме того, мы имеем возможность провести по способу наименьших квадратов параболу второй степени через надлежащее количество точек по обоим концам интервала. Производные соответствующих квадратичных функций в крайних точках дадут нам тогда значения $f'(b)$ и $f'(a)$, которыми можно воспользоваться для концевой поправки (23). С помощью этого приема можно значительно увеличить точность простого метода трапеций.

8. Численные примеры

Задача I. Для численной иллюстрации работы с различными формулами мы рассмотрим простой пример, который даст нам возможность проследить общий аналитический процесс, не наталкиваясь на большие технические затруднения. Пусть

$$y = e^x.$$

Положим $h=0,5$, примем за промежуток интегрирования интервал от $x=0$ до $x=4$.

Необходимые ординаты возьмем из таблицы значений функции $y=e^x$:

x	y
0	1
0,5	1,64872
1	2,71828
1,5	4,48169
2	7,38906
2,5	12,18249
3	20,08554
3,5	33,11545
4	54,59815

Теоретическое значение площади кривой в этой задаче равно

$$A = \int_a^b f(x) dx = \int_0^4 e^x dx = |e^x|_0^4 = e^4 - 1,$$

что дает

$$A = 53,59815.$$

По формуле трапеций находим:

$$\bar{A} = 0,5 \left(\frac{1}{2} + 1,64872 + 2,71828 + \dots + \frac{1}{2} 54,59815 \right) = 54,71015.$$

Погрешность этого результата составляет

$$53,59815 - 54,71015 = -1,11200.$$

Теоретическая оценка погрешности (6.5) дает

$$-\frac{h^2}{12} [y'(b) - y'(a)] = -\frac{1}{48} (54,59815 - 1) = -1,11663.$$

Теперь мы применим к этой же задаче формулу Симпсона:

$$\begin{aligned} \bar{A} = \frac{2}{3} 0,5 \left[\left(\frac{1}{2} + 2,71828 + 7,38906 + 20,08554 + \frac{1}{2} 54,59815 \right) + \right. \\ \left. + 2(1,64872 + 4,48169 + 12,18249 + 33,11545) \right] = 53,61622. \end{aligned}$$

Новая погрешность составляет $53,59815 - 53,61622 = -0,01807$.

Оценка погрешности (5.3) дает

$$-\frac{h^4}{180} [y'''(b) - y'''(a)] = -\frac{1}{2880} (54,598 - 1) = -0,01861,$$

тогда как из (5.4) $\bar{\eta} = -0,01816$.

Правило трапеций с концевой поправкой приводит к значению

$$54,71015 - 1,11663 = 53,59352.$$

Это значение несколько *меньше* точного значения, тогда как формула Симпсона дала значение немного *большее*, чем точное. Теперь ошибка равна

$$53,59815 - 53,59352 = 0,00463.$$

Эта погрешность составляет $1/4$ погрешности формулы Симпсона, т. е. равна

$$0,25 \cdot 0,01861 = 0,00465.$$

Во всех этих случаях точные и вычисленные по квадратурным формулам значения площади весьма близки. Теперь мы рассмотрим задачу, в которой условия складываются менее благоприятно.

Задача II. Функция, использованная в задаче I, была «гладкой», т. е. последовательные разности этой функции удовлетворительно убывали. Рассмотрим функцию, для которой разложение Стирлинга сходится гораздо менее удовлетворительно ввиду близости особой точки.

Рассмотрим функцию $y = \frac{1}{\sqrt{x}}$ в промежутке $(0,1; 1,7)$; равноотстоящие ординаты следуют друг за другом на расстоянии $h = 0,2$:

x	y
0,1	3,16228
0,3	1,82574
0,5	1,41421
0,7	1,19523
0,9	1,05409
1,1	0,95346
1,3	0,87706
1,5	0,81650
1,7	0,76696

Теоретическое значение интеграла:

$$\int_{0,1}^{0,7} x^{-1/2} dx = 2 \left| x^{1/2} \right|_{0,1}^{1,7} = 2 (\sqrt{1,7} - \sqrt{0,1}) = 2 \cdot 0,98756 = 1,97512.$$

По правилу трапеций получаем

$$\begin{aligned} \bar{A} &= 0,2 \left(\frac{1}{2} \cdot 3,16228 + 1,82574 + \dots + \frac{1}{2} 0,76696 \right) = \\ &= 0,2 \cdot 10,10092 = 2,02018 \quad [\text{погрешность: } - 0,045]; \end{aligned}$$

по формуле Симпсона

$$\bar{A} = \frac{2}{3} 0,2 \left[\left(\frac{1}{2} \cdot 3,16228 + 1,41421 + 1,05409 + \right. \right. \\ \left. \left. + 0,87706 + \frac{1}{2} \cdot 0,76696 \right) + 2 (1,82574 + 1,19523 + 0,95346 + \right. \\ \left. + 0,81605) \right] = 1,98558 \quad [\text{погрешность: } -0,010];$$

по правилу трапеций с концевой поправкой

$$\bar{A} = 2,02018 - \frac{0,04}{12} \left(-\frac{0,76696}{3,4} + \frac{3,16228}{0,2} \right) = \\ = 1,96823 \quad [\text{погрешность: } 0,0069].$$

В этом примере уточнение простого метода трапеций имеет гораздо меньшее влияние на результат, чем в предыдущем примере. Порядок величины погрешности остается одним и тем же во всех трех методах. Это обстоятельство объясняется тем, что влиянию высших степеней h противодействует сильное возрастание высших производных. Близость особой точки $x=0$ значительно уменьшает эффективность исчисления разностей, придавая гораздо больший вес старшим членам ряда Стирлинга,—это можно усмотреть, если вычислить оценки погрешности трех формул.

Оценка погрешности для формулы трапеций:

$$A - \bar{A} = -\frac{0,04}{12} \left(-\frac{0,767}{3,4} + \frac{3,162}{0,2} \right) = -0,051;$$

для формулы Симпсона [ср. (5.3)]:

$$A - \bar{A} = -\frac{0,0016}{180} \left(-\frac{15}{8} \frac{0,767}{(1,7)^3} + \frac{15}{8} \frac{3,162}{(0,1)^3} \right) = -0,053;$$

для скорректированной формулы трапеций:

$$A - \bar{A} = \frac{0,053}{4} = 0,013.$$

Ненадежность оценки погрешности в случае формулы Симпсона обусловлена негладкостью функции; это имеет своим следствием то, что третья производная и третья центральная разность в окрестности нижнего предела далеки друг от друга. Формула (5.4) дает более надежный результат. Применяя ее, получаем $\bar{\eta} = -0,0139$, что превышает точную величину ошибки $\eta = -0,010$, но не чрезмерно.

9. Приближение полиномами высших степеней

В формуле Симпсона мы заканчивали ряд Стирлинга *квадратным членом*. Мы можем, очевидно, идти дальше и закончить ряд членом *более высокой степени*. Соответственно этому возрастет число необходимых для расчета полос. Так как мы не хотим терять большого

преимущества, возникающего от симметричных пределов, то следующим шагом после пределов ± 1 будут пределы ± 2 , при которых в расчет войдут четыре соседние полосы. Это означает, что мы используем пять последовательных данных:

$$y_0, y_1, y_2, y_3, y_4, \quad (6-9.1)$$

с помощью которых можно образовать центральные разности вплоть до четвертого порядка. Интегрируя в пределах от -2 до $+2$, получаем

$$\begin{aligned} \bar{A} &= \int_{-2}^{+2} f(2+t) dt = \left| f(2)t + \frac{\delta^2 f(2)}{2} \frac{t^3}{3} + \frac{\delta^4 f(2)}{24} \left(\frac{t^5}{5} - \frac{t^3}{3} \right) \right|_{-2}^{+2} = \\ &= 4f(2) + \frac{8}{3} \delta^2 f(2) + \frac{14}{45} \delta^4 f(2) = \\ &= \frac{2}{45} [90f(2) + 60\delta^2 f(2) + 7\delta^4 f(2)] = \\ &= \frac{2}{45} [7y_0 + 32y_1 + 12y_2 + 32y_3 + 7y_4]. \quad (6-9.2) \end{aligned}$$

Если ширина полосы есть h , то правую часть формулы следует умножить на h . Таким образом, окончательная пятиточечная формула имеет вид:

$$\bar{A} = \frac{2h}{45} (7y_0 + 32y_1 + 12y_2 + 32y_3 + 7y_4). \quad (6-9.3)$$

Погрешность этого приближения может быть снова оценена по первому опущенному члену в разложении Стирлинга:

$$\begin{aligned} \bar{\eta} &= \frac{h\delta^6 f(2h)}{6!} \int_{-2}^{+2} t^2 (t^2 - 1) (t^2 - 4) dt = \\ &= \frac{h}{720} \delta^6 f(2h) \left| \frac{t^7}{7} - 5\frac{t^5}{5} + 4\frac{t^3}{3} \right|_{-2}^{+2} = -\frac{8h}{945} \delta^6 f(2h). \quad (6-9.4) \end{aligned}$$

Если шестой разностный коэффициент заменить шестой производной, то оценка погрешности четырехполосной формулы примет вид

$$\bar{\eta} = -\frac{8h^7}{945} f^{(6)}(2h). \quad (6-9.5)$$

В нашем предыдущем численном примере вся область была поделена на 8 полос равной ширины. Применяя формулу Симпсона, мы сгруппировали эти полосы в 4 двойные полосы. Теперь мы сгруппируем их в 2 четырехполосные группы. В силу того, что есть крайняя правая ордината в первой группе и одновременно крайняя левая ордината второй группы, веса последовательных ординат будут

$$\frac{2h}{45} (7, 32, 12, 32, 14, 32, 12, 32, 7). \quad (6-9.6)$$

Беря табличные ординаты, указанные в первом примере предыдущего параграфа, с этими весами получаем

$$\bar{A} = \frac{2 \cdot 0,5}{45} 2411,98706 = 53,59971.$$

Погрешность этого результата составляет

$$A - \bar{A} = 53,59815 - 53,59971 = -0,00156.$$

Оценивая погрешность на основе формулы (5), нужно помнить, что мы пользовались *двумя* группами полос. Соответственно этому второй множитель надо взять в точках $2h=1$ и $6h=3$; вычислив их сумму, получим

$$f^{(6)}(1) + f^{(6)}(3) = 2,72 + 20,09 = 22,81;$$

это дает

$$A - \bar{A} = -\frac{8 \cdot (0,5)^7}{945} 22,81 = -0,00151.$$

Точность этой оценки достаточно высока.

Отметим, что приближение четвертой степени уменьшило погрешность в 12 раз по сравнению с приближением второй степени по формуле Симпсона. Этот выигрыш обусловлен более высокой степенью h , которой здесь не противодействует чрезмерно большое возрастание старших производных.

Совершенно другое положение оказывается во втором примере, где высшие производные очень быстро возрастают при малых значениях x . Если в этой задаче взять веса (6), то получим

$$\begin{aligned} \bar{A} = \frac{0,4}{45} [& 14(1,581139 + 1,054093 + 0,383482) + \\ & + 32(1,825742 + 1,195229 + 0,953463 + 0,816497) + \\ & + 12(1,414214 + 0,877058)] = 1,982818. \end{aligned}$$

Ошибка теперь составляет $-0,0077$, она лишь немногим меньше прежней ошибки $-0,010$. Здесь, в противоположность предыдущему примеру, нам не удастся получить надежную оценку погрешности. Причиной является то, что оценка по формуле (4) основывалась на сходимости разложения Стирлинга, а это в свою очередь возможно лишь при гладком характере разностей старших порядков. В последнем нашем примере близость особой точки $x=0$ мешает образованию шестой разности. Поэтому мы не можем ожидать хорошего результата от применения формулы (5). Далее мы увидим, что метод § 12, который не требует знания производных выше первого порядка, действует и здесь удовлетворительно. Его применение в настоящей задаче дает оценку погрешности $\bar{\eta} = -0,0127$, которая несколько превышает, но не чрезмерно, истинную ошибку $\eta = -0,0077$.

Вообще, можно сказать, что аппроксимирование с использованием интерполяционного полинома высокой степени полезно только в том случае, когда одновременно h выбрано достаточно малым. Можно значительно выиграть в точности, если при достаточно малом h оперировать полиномами не слишком низкой степени. Было бы, однако, ошибкой думать, что мы всегда получим большую точность, аппроксимируя *всю* область одним полиномом степени N , если всех ординат имеется $N+1$. Этому препятствует, вообще говоря, расходящийся характер равноотстоящего приближения полиномами (см. V, 15). На практике превосходная точность формулы Симпсона обычно бывает достаточной, так как она соединяет сравнительно высокую степень h , а именно четвертую, с производной сравнительно низкого порядка, а именно третьего [ср. (5.3)]. В случае гладких функций заслуживает внимания четырехполосная формула (3). Иногда бывает полезна и следующая шестиполосная формула, известная, как «правило Уэддла»:

$$\bar{A} = \frac{3h}{10} (y_0 + 5y_1 + y_2 + 6y_3 + y_4 + 5y_5 + y_6). \quad (6-9.7)$$

Эта формула оперирует с параболой шестой степени, но для упрощения в ней немного видоизменен вес шестой разности.

Весьма полезным является также простое правило трапеции с добавлением концевой поправки [ср. (7.21)]. Погрешность этой формулы составляет $1/4$ погрешности формулы Симпсона. Она может применяться для проверки и особенно целесообразна, если заданные ординаты являются результатом наблюдений, а не вычислений.

10. Гауссов метод квадратуры

Выдающийся математик Гаусс внес совершенно новую и исключительно остроумную идею в обычную теорию квадратуры. Эта идея в своем дальнейшем развитии стала основополагающей во многих областях практического анализа.

Допустим, что некоторая интегрируемая функция $y = f(x)$ задана не в каждой точке непрерывного промежутка переменной x , а лишь в специально выбранных точках $x_1, x_2, \dots, x_n$, которые лежат внутри данного промежутка. Так как мы имеем в виду лишь *конечный* промежуток, то мы сразу же можем *нормировать* его. Поместим начало отсчета переменной в середину рассматриваемого промежутка и выберем такой масштабный множитель, который переводит обе конечные точки промежутка в точки $x = \pm 1$. Мы, следовательно, рассматриваем промежуток

$$-1 \leq x \leq +1 \quad (6-10.1)$$

и принимаем, что точки $x_1, x_2, \dots, x_n$, в которых задана

функция $y = f(x)$, принадлежат этому промежутку. Ординат

$$\left. \begin{aligned} y_1 &= f(x_1), \\ y_2 &= f(x_2), \\ &\dots\dots\dots \\ y_n &= f(x_n), \end{aligned} \right\} \quad (6-10.2)$$

вообще говоря, недостаточно для определения функции $f(x)$, независимо от того, как велико будет число n . Но мы можем постараться проинтерполировать $f(x)$ для промежуточных точек. С этой целью воспользуемся степенными функциями от x . Мы можем найти такой полином $p_{n-1}(x)$ степени $n-1$, который обладает тем свойством, что он принимает заданные значения y_k в данных точках x_k .

Обычно в исчислении конечных разностей принимается, что заданные точки $x = x_k$ выбраны *равноотстоящими*. Идея Гаусса состоит в том, что мы могли бы, возможно, получить бо́льшую точность при том же числе ординат, если не фиксировать заранее их положение, а использовать расположение задаваемых точек таким образом, чтобы получить наилучшие результаты.

На этом пути Гауссу удалось получить не только формулу квадратуры исключительной точности, но также процесс, свободный от опасностей равноотстоящей интерполяции полиномами, хотя эти опасности еще не были известны в его время.

Допустим, что точки интерполяции $x = x_k$ остаются совершенно свободными, и мы хотим найти полином $u = p_{n-1}(x)$, который принимает в этих точках заданные значения $y_1, y_2, \dots, y_n$. Формула, решающая эту задачу, известна как «интерполяционная формула Лагранжа»¹⁾. Она основана на построении фундаментального полинома

$$F_n(x) = (x - x_1)(x - x_2) \dots (x - x_n) \quad (6-10.3)$$

и последовательного деления его на каждый из n корневых двучленов $(x - x_1), \dots, (x - x_n)$. Мы получаем, таким образом, ряд полиномов

$$Q_i(x) = \frac{F_n(x)}{F'_n(x_i)(x - x_i)} \quad (i = 1, 2, \dots, n), \quad (6-10.4)$$

которые обладают следующими свойствами: $Q_i(x)$ обращаются в нуль во всех точках $x = x_k$, кроме точки $x = x_i$, в которой $Q_i(x)$ равно 1. Если ввести «символ Кронекера» δ_{ik} , который принимает значение 1 при $i = k$ и нуль при $i \neq k$, то, очевидно,

$$Q_i(x_k) = \delta_{ik}. \quad (6-10.5)$$

Но тогда, как легко видеть, полином

$$p_{n-1}(x) = y_1 Q_1(x) + y_2 Q_2(x) + \dots + y_n Q_n(x) \quad (6-10.6)$$

¹⁾ См. {8}, {11}.

удовлетворяет поставленному условию: он принимает в точках $x = x_k$ заданные значения $y = y_k$ ($k = 1, 2, \dots, n$). Единственность полинома $p_{n-1}(x)$ следует из того факта, что разность между $p_{n-1}(x)$ и гипотетическим вторым полиномом $\bar{p}_{n-1}(x)$ обращалась бы в нуль во всех точках $x = x_k$. Но разность $p_{n-1}(x) - \bar{p}_{n-1}(x)$ есть снова полином степени $n - 1$, а такой полином не может иметь больше, чем $n - 1$ корней, не обращаясь тождественно в нуль; это означает, что $\bar{p}_{n-1}(x) = p_{n-1}(x)$.

Теперь, если считать $p_{n-1}(x)$ достаточно близким приближением функции $y = f(x)$, можно получить парактически*) площадь под неизвестной кривой $f(x)$, вычислив

$$\bar{A} = \int_{-1}^{+1} p_{n-1}(x) dx = \sum_{k=1}^n y_k \int_{-1}^{+1} Q_k dx. \quad (6-10.7)$$

Для некоторого заданного распределения точек $x = x_i$ полиномы $Q_k(x)$ однозначно определены, и поэтому определенные интегралы

$$\int_{-1}^{+1} Q_k(x) dx = w_k \quad (6-10.8)$$

имеют некоторые определенные численные значения, для которых могут быть составлены таблицы. Эти значения совершенно не зависят от природы функции $y = f(x)$, площадь которой нас интересует.

Остроумный метод квадратуры Гаусса можно теперь ввести следующим образом. Добавим к ранее заданным точкам дополнительную точку

$$x = x_{n+1},$$

не меняя предыдущих точек x_i . Вводя дополнительный корневой двучлен $x - x_{n+1}$, образуем дополнительный полином $Q_{n+1}(x)$. Из определения (4) для $Q_i(x)$ следует, что этот полином $Q_{n+1}(x)$ будет пропорционален полиному $F_n(x)$, так как новый множитель $(x - x_{n+1})$ сокращается. Следовательно, весовой множитель w_{n+1} , на который надо умножить новую ординату y_{n+1} , будет пропорционален определенному интегралу

$$\int_{-1}^{+1} F_n(x) dx. \quad (6-10.9)$$

Точно так же, если ввести m новых точек

$$x = x_{n+1}, x_{n+2}, \dots, x_{n+m} \quad (6-10.10)$$

*) Смысл этого термина пояснен во Введении.

вместе с их ординатами, то соответствующие им веса $w_{n+1}, w_{n+2}, \dots, w_{n+m}$ определяются интегралами вида

$$w_{n+i} = \int_{-1}^{+1} F_n(x) \mathcal{G}_{m-1}^i(x) dx, \quad (6-10.11)$$

где $\mathcal{G}_{m-1}^i(x)$ суть некоторые полиномы степени $m-1$. В силу того факта, что произвольный полином $\mathcal{G}_{m-1}(x)$ есть линейная суперпозиция степенных функций $x^0, x^1, x^2, \dots, x^{m-1}$, все эти веса *автоматически обратятся в нуль*, если $F_n(x)$ будет удовлетворять следующим интегральным условиям:

$$\int_{-1}^{+1} F_n(x) dx = 0, \quad \dots, \quad \int_{-1}^{+1} F_n(x) x^{m-1} dx = 0. \quad (6-10.12)$$

В действительности мы можем дойти до $m=n$ в наших требованиях выполнения интегральных условий

$$\int_{-1}^{+1} F_n(x) x^\alpha dx = 0 \quad (\alpha = 0, 1, 2, \dots, n-1). \quad (6-10.13)$$

В результате мы можем свободно добавить к нашим первоначально заданным n точкам любые другие n точек, и все-таки ни одна из новых ординат ничего не меняет в полученном ранее результате*). Полученный результат будет таков, как будто мы оперировали с $2n$ ординатами, а *фактически* мы пользуемся лишь n ординатами, так как все дополнительные ординаты ничего не прибавляют к вычисляемой площади. При этой процедуре мы экономим n членов в сумме

$$\bar{A} = \sum_{k=1}^{2n} y_k w_k.$$

Рассуждения, приведенные автором в доказательство этого утверждения, нельзя считать достаточными. Действительно, при задании новых дополнительных точек $x_{n+1}, x_{n+2}, \dots, x_{2n}$ не только добавляются *новые* полиномы Q_{n+m} ($m=1, 2, \dots, n$), но и *прежние* полиномы Q_i ($i=1, 2, \dots, n$) изменяются: каждая новая точка x_{n+m} вводит в Q_i добавочный множитель

$\frac{x - x_{n+m}}{x_i - x_{n+m}}$. Таким образом, введение новых m точек $x_{n+1}, x_{n+2}, \dots, x_{n+m}$ превращает прежний полином Q_i в полином

$$Q_i^* = Q_i \frac{x - x_{n+1}}{x_i - x_{n+1}} \cdot \frac{x - x_{n+2}}{x_i - x_{n+2}} \cdot \dots \cdot \frac{x - x_{n+m}}{x_i - x_{n+m}}. \quad (a)$$

Справедливость приведенного в тексте положения вытекает из того, что эти видоизмененные полиномы Q_i^* обладают следующими свойствами:

Но еще более важен тот факт, что мы не должны даже *знать* дополнительных ординат $y_{n+1}, y_{n+2}, \dots, y_{2n}$.

Сумма

$$\bar{A} = \sum_{k=1}^n y_k w_k \quad (6-10.14)$$

дает площадь с помощью n ординат с такой точностью, как если бы мы пользовались $2n$ ординатами.

Интегральные условия типа (13) называются «условиями ортогональности». Мы говорим, что полином $F_n(x)$ «ортогонален» степенным функциям $1, x, x^2, \dots, x^{n-1}$. Мы встречали такие условия раньше при рассмотрении «систем ортогональных функций» (см. V, 16). Мы исследовали полиномы Якоби (см. V, 20), которые обладают тем свойством, что они ортогональны ко всем степеням x , меньшим степени полинома, как раз в смысле условий (13). Однако в общем случае условие ортогональности включает еще множитель веса $\rho(x)$ под знаком интеграла. Только в особом случае «полиномов Лежандра» [ср. (5-20.11)] этот множитель веса оказывается равным 1 и, таким образом, взвешенная ортогональность превращается в обыкновенную ортогональность. Тем самым решается вопрос о выборе функции $F_n(x)$; метод Гаусса требует отождествления $F_n(x)$ с n -м полиномом Лежандра $P_n(x)$; нули этих полиномов дают нам точки, в которых должны быть заданы значения функции $f(x)$. Имеются весьма точные таблицы этих нулей вместе с численными значениями коэффициентов w_i , которые вычисляются по формуле (8) (ср. также § 13)¹⁾.

$$1^\circ. \quad Q_i^*(x_k) = \delta_{ik} \quad (k = 1, 2, \dots, 2n).$$

$$2^\circ. \quad \int_{-1}^{+1} Q_i^*(x) dx = \int_{-1}^{+1} Q_i(x) dx = w_i.$$

Первое свойство легко усмотреть непосредственно из соотношения (α). Для доказательства 2° заметим, что

$$\frac{x - x_{n+k}}{x_i - x_{n+k}} = 1 + \frac{x - x_i}{x_i - x_{n+k}}.$$

Отсюда заключаем, что произведение дополнительных множителей в правой части равенства (α) можно представить в виде $1 - \mathcal{G}_{m-1}^i$, где $\mathcal{G}_{m-1}^i$ — полином $(m-1)$ -й степени. Но тогда, в силу условий (10.13), выполняется равенство 2°.

Доказанные свойства 1° и 2° и составляют точное содержание утверждения, что новые ординаты «ничего не меняют в полученном ранее результате».

¹⁾ См. [2]. Труд вычисления коэффициентов w_i наполовину сокращается в силу симметрических свойств полиномов Лежандра. Корни их $\pm \xi_k$ все парные; веса, соответствующие двум таким точкам, равны между собой.

11. Численный пример

Мы применим метод Гаусса к тем же численным примерам, которые были рассмотрены в § 8. В первом примере мы имеем очень гладкую, а во втором — очень негладкую функцию. Поэтому скорость сходимости квадратурных формул при большом числе точек n весьма различна в этих двух случаях. В наших предыдущих процессах мы пользовались девятью равноотстоящими ординатами. Теперь мы заменим их только пятью неравноотстоящими ординатами. Так как таблицы гауссовых нулей рассчитаны для интервала $(-1, 1)$, необходимо приспособить произвольный интервал к этому нормированию. Это мы осуществляем с помощью преобразования

$$x = \frac{b+a}{2} + \frac{b-a}{2} \xi, \quad (6-11.1)$$

$$\int_a^b f(x) dx = \frac{b-a}{2} \int_{-1}^{+1} f(\xi) d\xi. \quad (6-11.2)$$

В первом нашем примере $a=0$, $b=4$; нужно, следовательно, найти ординаты в точках

$$x_k = 2 + 2\xi_k = 2(1 + \xi_k). \quad (6-11.3)$$

При выборе $n=5$ мы получаем пять нулей:

$$\begin{aligned} &2(1 \pm 0,906179846) \\ &2(1 \pm 0,538469310) \\ &2 \end{aligned}$$

Таким образом, точками интерполяции вместе с соответствующими весами w_k будут (с точностью до 8 десятичных знаков)

$$\begin{array}{ll} x_1 = 0,18764031, & w_1 = 0,47385377, \\ x_2 = 0,92306138, & w_2 = 0,95725734, \\ x_3 = 2, & w_3 = 1,13777778, \\ x_4 = 3,07693862, & w_4 = 0,95725734, \\ x_5 = 3,81235969, & w_5 = 0,47385377. \end{array} \quad (6-11.4)$$

Эти веса были получены умножением гауссовых весов на $(b-a)/2$ *). Соответствующими ординатами будут

$$\begin{aligned} y_1 &= 1,20639950, \\ y_2 &= 2,51698405, \\ y_3 &= 7,38905610, \\ y_4 &= 21,69189349, \\ y_5 &= 45,25710562. \end{aligned}$$

*) Ординаты функции $y=e^x$ в этих точках можно найти в [4], производя надлежащие интерполяции.

Умножая эти ординаты на их веса из таблицы (4) и суммируя, получаем

$$\bar{A} = 53,59813663$$

при истинном значении

$$A = e^4 - 1 = 53,59815003.$$

Ошибка приближения равна

$$\eta = 0,0000134.$$

Найденное выше приближение (9,6) с помощью девяти ординат дало гораздо бóльшую ошибку:

$$\eta = -0,0016.$$

Таким образом, метод Гаусса дает удивительную точность. Несмотря на то, что мы оперировали с пятью ординатами вместо девяти, погрешность уменьшилась более чем в 100 раз. Создается впечатление, что приближение с неравными интервалами дает даже *большой* эффект, чем тот, на который можно было рассчитывать. Это приближение оказывает благоприятное влияние на уменьшение ошибки в дополнение к тому, что избавляет от некоторого количества ординат. И это действительно так. Равноотстоящее интерполирование вообще не является хорошо сходящимся процессом, а для функций, имеющих особые точки внутри единичного круга (хотя они могут быть совершенно гладкими между -1 и $+1$), сама сходимость, вообще говоря, не может быть даже гарантирована (ср. V, 15). Гауссов процесс квадратуры использует в качестве точек интерполяции нули полиномов Лежандра, т. е. нули *ортogonalного* множества полиномов. Сходимость этого процесса гарантируется общими свойствами ортогональных разложений (ср. V, 21).

Метод квадратур Гаусса, таким образом, превосходит обычные «равноотстоящие методы» в двояком отношении. Во-первых, эффективность n неравноотстоящих ординат такая же, как эффективность $2n$ равноотстоящих ординат. Во-вторых, интерполяция полиномами Лежандра сходится гораздо лучше, нежели интерполяция полиномами Лагранжа.

Во втором примере § 8 слабо сходится даже квадратура Гаусса. Но и здесь мы снова можем показать силу метода Гаусса, сэкономив число ординат. На этот раз мы воспользуемся только четырьмя ординатами вместо первоначально взятых девяти. Пределами теперь будут

$$a = 0,1; \quad b = 1,7,$$

следовательно,

$$x_k = 0,9 \pm 0,8\xi_k;$$

четыре точки интерполяции суть

$$0,9 \pm 0,8 \cdot 0,33998104,$$

$$0,9 \pm 0,8 \cdot 0,86113631,$$

и мы можем составить таблицу

$$\begin{array}{ll} x_1 = 0,21109095, & w_1 = 0,34785484, \\ x_2 = 0,62801516, & w_2 = 0,65214515, \\ x_3 = 1,17198483, & w_3 = 0,65214515, \\ x_4 = 1,58890905, & w_4 = 0,34785484. \end{array}$$

На этот раз мы выписали веса из общей таблицы, оставив их неизменными, так как численно проще умножить полученный результат на $\frac{b-a}{2} = 0,8$, чем умножать каждый вес на этот множитель.

Ординаты y_k функции $y = x^{-\frac{1}{2}}$ в точках интерполяции теперь будут

$$\begin{array}{l} y_1 = 2,17653268, \\ y_2 = 1,26187092, \\ y_3 = 0,92371714, \\ y_4 = 0,79332380. \end{array}$$

Умножение на соответствующие веса и суммирование дают

$$2,45839962,$$

следовательно,

$$\bar{A} = 0,8 \cdot 2,45839962 = 1,96671970.$$

Точное значение площади здесь равно

$$A = 1,97512,$$

что дает ошибку

$$\eta = 0,0084.$$

Это лишь немногим больше, чем ошибка, полученная в § 9 при использовании девяти ординат:

$$\eta = -0,0077.$$

В первом случае мы делили промежуток интегрирования на две полосы и в каждой полосе пользовались аппроксимирующим полиномом четвертой степени. Теперь весь интервал как будто аппроксимирован полиномом третьей степени. Фактически был использован полином *седьмой степени*, ибо каждая наша точка в действительности считалась *двойной точкой*. Мы прокладываем параболу седьмой степени через восемь специально выбранных точек. Точки интерполяции не просто четыре нуля полинома $P_4(\xi)$. Они фактически — четыре *пары* точек, но точки каждой пары очень близки друг к другу и в пределе сливаются в одну точку. Четыре нуля полинома $P_4(\xi)$ в действительности заменяют *восемь* нулей функции $P_4^2(\xi)$.

12. Погрешность квадратуры Гаусса

Чем более эффективен какой-нибудь метод парэктического анализа, тем труднее обычно получить удовлетворительную оценку достигаемой при этом точности. В случае квадратуры Гаусса нелегко оценить погрешность. Традиционная формула оценки погрешности в этом случае требует знания $2n$ -й производной функции $f(x)$ на всем интервале интегрирования¹⁾. Формула эта такова:

$$\eta = \left[\frac{(n!)^2}{(2n)!} \right]^2 \frac{2^{2n+1}}{2n+1} \frac{f^{(2n)}(\theta)}{(2n)!}, \quad (6-12.1)$$

где θ — некоторое число, заключенное между -1 и $+1$. Оценка эта имеет различные недостатки. С теоретической точки зрения можно возражать против предпосылки, что $f(x)$ обладает $2n$ производными в промежутке интегрирования, тогда как квадратура Гаусса сходится к соответствующей величине даже в случае неаналитических функций, таких, как, например, $\sqrt{|x|}$, первая производная которой становится бесконечно большой при $x=0$ и которая все же отлично поддается квадратуре Гаусса. С практической точки зрения обычно трудно (за исключением простых случаев) вычислить $2n$ -ю производную, даже если известна аналитическая форма функции. С другой стороны, часто $f(x)$ задается лишь таблично, и аналитическое выражение этой функции неизвестно.

Нижеследующая схема свободна от этих недостатков. Если применить квадратуру Гаусса к $f'(x)$, то, очевидно, результат должен быть

$$\int_{-1}^{+1} f'(x) dx = f(1) - f(-1).$$

Следовательно, в этом случае мы можем непосредственно проверить ошибку квадратуры Гаусса. Чтобы надлежащим образом использовать эту идею, нужно, однако, учесть, что квадратура в пределах от -1 до $+1$ включает лишь *четную* часть функции, а именно $f(x) + f(-x)$, тогда как квадратура производной включала бы совершенно другую функцию, а именно *нечетную* часть, которая равна $f(x) - f(-x)$. Мы обойдем это затруднение, беря производную функции $xf(x)$, которая обладает симметрией такого же характера, что и сама функция $f(x)$:

$$\int_{-1}^{+1} [xf(x)]' dx = f(1) + f(-1). \quad (6-12.2)$$

Так как $(xf)' = xf' + f$, то численная процедура, как видим,

¹⁾ Ср. [2], стр. 740.

получается простой. Умножаем прежние веса w_i на $\xi_i f'(\xi_i)$ вместо $f(\xi_i)$. Тогда погрешность новой квадратуры дается формулой

$$\eta' = f(1) + f(-1) - \bar{A} - \sum_{\alpha=1}^n w_{\alpha} \xi_{\alpha} f'(\xi_{\alpha}). \quad (6-12.3)$$

В случае произвольных пределов a и b [ср. (11.1)]:

$$\eta' = \frac{b-a}{2} [f(b) + f(a)] - \bar{A} - \left(\frac{b-a}{2}\right)^2 \sum_{\alpha=1}^n w_{\alpha} \xi_{\alpha} f'(x_{\alpha}). \quad (6-12.4)$$

Более тщательное исследование погрешности квадратуры Гаусса обнаруживает, что для функций, не слишком негладких между -1 и $+1$, величина θ в формуле (1) близка к нулю. Но тогда последний множитель в этой формуле практически равен коэффициенту при ξ^{2n} разложения Тейлора в окрестности начала. Далее, разложение функции $(\xi f')$ тождественно разложению первоначальной функции, за исключением того, что a_{2n} умножается на $2n + 1$. При этих условиях мы получаем оценку погрешности в следующем виде:

$$\bar{\eta} = \frac{1}{2n+1} \eta'. \quad (6-12.5)$$

Большим преимуществом этой оценки является то, что она требует знания лишь *первой* производной от $f(x)$ вместо $2n$ -й производной, входящей в традиционную формулу. Можно было бы подумать, что $2n$ -я производная от $f(x)$ необходима, если учесть тот факт, что квадратура точна для всякого полинома степени, меньшей чем $2n$. Для того чтобы исключить такой полином, мы должны дифференцировать $2n$ раз. Но формула (5), хотя мы продифференцировали только один раз, ничего не теряет в этом отношении. Если $f(x)$ есть полином степени меньше $2n$, то и функция $(xf)'$ есть снова полином такого же типа. Следовательно, вторая квадратура дает ошибку нуль, и оценка (5) также дает нуль.

Если функция $f(x)$ не столь гладкая, как это предполагалось в вышеприведенном рассуждении, то можно принять, что высшие производные будут возрастать скорее, чем убывает негладкость функции. Мы поэтому можем считать, что оценка по формуле (5) будет слишком груба. Мы, однако, не можем быть уверены в оценке (5) и в том случае, когда $f'(x)$ меняет знак в заданном интервале¹⁾.

Мы применим теперь этот метод оценки погрешности квадратуры Гаусса к двум численным примерам § 11. В первом из них получаем

$$\eta' = 2 \cdot 55,59815003 - 53,59813663 - 57,59801484 = 0,0001487.$$

Деля на $2n + 1 = 11$, находим

$$\bar{\eta} = 0,0000135,$$

¹⁾ Точные пределы надежности этого приема до сих пор еще не установлены.

раньше (ср. IV, 23) и получили простой численный алгоритм для ее решения [ср. (IV-23.11,12,13)]. Этот алгоритм применим к нашей задаче и дает более простой метод вычисления w_i , чем непосредственное вычисление определенных интегралов (10.8).

14. Квадратура Гаусса с округленными узлами

С практической точки зрения квадратура Гаусса страдает одним серьезным недостатком. Функция должна быть вычислена в иррациональных точках. Это требует сложной и довольно громоздкой интерполяции. В случае табличных функций для получения n интерполированных ординат зачастую потребуется гораздо больше труда, чем для того, чтобы просчитать непосредственно $2n$ равноотстоящих ординат. По этой причине метод Гаусса обыкновенно применяется лишь в том случае, когда из-за отсутствия каких-либо таблиц для каждого значения y_k требуется отдельное независимое вычисление.

Этот недостаток метода Гаусса можно исправить следующим видоизменением первоначальной схемы. Мы *округляем* гауссовы нули (т. е. узловые точки) до небольшого числа десятичных знаков, скажем до двух или трех, и вычисляем веса, соответствующие этим смещенным нулям. Тогда процесс интерполяции очень упрощается или даже совсем устраняется, если мы располагаем таблицами функции $f(x)$, построенными с шагом 0,01 или 0,001 для аргумента x . Правда, полной точности процесса Гаусса при этом нельзя достичь, но точность все еще велика. В действительности при некоторой схеме дополнительной корректировки, которую мы рассмотрим в следующем параграфе, можно *полностью сохранить* точность метода Гаусса.

Коэффициенты w_i , соответствующие смещенным узлам, можно вычислить по общему методу, указанному в § 13. Строим сначала фундаментальный полином $F_n(\xi)$ с помощью корневых двучленов. Выберем, например, $n=5$. Тогда пять гауссовых узлов суть

$$\xi = \pm 0,53846, \quad 0, \quad \pm 0,90617.$$

Округлим их до двух десятичных знаков:

$$\xi = \pm 0,54, \quad 0, \quad \pm 0,91$$

и построим фундаментальный полином для этих узлов:

$$F_5(\xi) = \xi(\xi^2 - 0,54^2)(\xi^2 - 0,91^2) = \xi^5 - 1,1197\xi^3 + 0,24147396\xi.$$

Сравнение с пятым полиномом Лежандра

$$P_5(\xi) = \frac{1}{8}(63\xi^5 - 70\xi^3 + 15\xi)$$

требует умножения на 63 и введения множителя $\frac{1}{8}$ перед скобками:

$$\frac{1}{8}(63\xi^5 - 70,5411\xi^3 + 15,21285948\xi).$$

Отсюда видно, что оба полинома имеют почти одинаковые коэффициенты. Величины y_k из уравнений (4-23.8) в данном случае превращаются в определенные интегралы (13.1) при $\rho(x) = 1$:

$$\int_{-1}^{+1} x^k dx = \begin{cases} \frac{2}{k+1} & (k = 0, 2, 4 \dots), \\ 0 & (k = 1, 3, 5 \dots), \end{cases}$$

откуда выводится обратный полином

$$2\xi^{-1} + \frac{2}{3}\xi^{-3} + \frac{2}{5}\xi^{-5}.$$

На него надо умножить $F_5(\xi)$, согласно схеме (4-23.11),

$$\begin{array}{r} 0,4829479200, \quad 0, \quad -2,2394000000, \quad 0, \quad 2, \\ -0,7464666667, \quad 0, \quad 0,6666666667, \\ 0,4 \\ \hline 0,1364812533, \quad 0, \quad -1,5727333333, \quad 0, \quad 2. \end{array}$$

Отсюда

$$G_4(\xi) = 0,1364812533 - 1,5727333333\xi^2 + 2\xi^4$$

и

$$\frac{G_4(\xi)}{F'_5(\xi)} = \frac{0,1364812533 - 1,5727333333\xi^2 + 2\xi^4}{0,24147396 - 3,3591\xi^2 + 5\xi^4}$$

Подстановка значений $\xi = \pm 0,91; \pm 0,54; 0$ дает пять весов w_i квадратурной формулы¹⁾

$$\begin{array}{ll} \xi = \pm 0,91, & w = 0,231387878, \\ \pm 0,54, & 0,48601767, \\ 0, & 0,565200708. \end{array}$$

Эти веса мы применяем к примеру $y = e^x$, рассмотренному ранее (ср. § 11) с помощью точных узлов Гаусса. Получаем новые ординаты

$$\begin{array}{ll} x = 0,18 & y = 1,19721736, \\ 0,92, & 2,50929039, \\ 2 & 7,38905610, \\ 3,08, & 21,75840240, \\ 3,82, & 45,60420832. \end{array}$$

Веса должны быть умножены на 2, так как интервал переменной x удвоен. Сумма взвешенных ординат равна

$$\bar{A} = 53,59910051,$$

¹⁾ Систематическая таблица для операций с округленными узлами составлена в Проектном бюро математических таблиц (New York). Она перепечатана в приложении к настоящей книге (см. табл. XIV).

в то время как точное значение

$$A = 53,59815003.$$

Погрешность приближения составляет таким образом,

$$\eta = -0,00095048.$$

По сравнению с гауссовой погрешностью эта погрешность возросла в 71 раз, что свидетельствует о большой чувствительности метода Гаусса даже к небольшим смещениям узлов. Тем не менее точность результата все еще значительна.

Оценка погрешности на основе метода, разобранный в § 12, опять действует удовлетворительно. Мы получаем теперь

$$\eta' = 111,19630006 - 53,59910051 - 57,6093084 = -0,012109.$$

Деление на 11 дает

$$\bar{\eta} = -0,00110,$$

что только немногим больше указанной выше истинной ошибки.

15. Применение двукратных корней

В § 11 мы отметили, что большая эффективность метода Гаусса может быть объяснена тем обстоятельством, что фундаментальным полиномом является не $P_n(\xi)$, а фактически $P_n^2(\xi)$. Каждую точку интерполяции можно считать двойной точкой, так как по свойству квадратуры Гаусса n точек можно свободно добавить к точкам интерполяции, ничего не меняя, ибо новые точки входят в квадратурную формулу с весом нуль. Допустим, что мы выбрали произвольную совокупность n точек внутри интервала, но принимаем каждую точку за двукратный корень полинома $F_n(\xi)$. Так как двойные корни эквивалентны двум близким точкам, в пределе сливающимся в одну, то интерполяция с помощью n двойных точек означает, что в каждой точке интерполяции задаются значения функции $f(x_k)$ и ее производной $f'(x_k)$. Но гауссовы точки выбираются таким образом, что веса производных должны равняться нулю. Квадратурная формула, таким образом, сводится к n членам (вместо $2n$).

Если теперь слегка сместить гауссовы нули, то веса производных уже больше не будут равны нулю, но они останутся малыми. Поэтому не возникает необходимости получения производных с большой точностью. Если функция задается таблично через достаточно малые интервалы, простой разностный коэффициент между двумя соседними табличными значениями может заменить производную. Этим путем мы можем округлить гауссовы нули до удобных чисел, избежав таким образом неудобств интерполяции и все же сохранив полностью точность метода Гаусса.

В качестве численного примера мы снова обращаемся к уже рассмотренной показательной функции, используя пять гауссовых точек (ср. § 11). Округляем эти точки до $\xi = \pm 0,54$, $\xi = \pm 0,90$ (жертвуя при этом несколько лучшим значением 0,91 в пользу более удобного). Коэффициенты w_i и w'_i опять найдем по численной схеме предшествующего параграфа, повысив, однако, степень полинома $F_n(\xi)$ до 10 (путем возведения в квадрат). Соответственно полином $\mathcal{G}_{n-1}(\xi)$ будет иметь степень 9, но фактически мы придем к полиному четвертой степени относительно ξ^2 , так как все нечетные степени ξ выпадают. В результате мы получаем пять весов w_i ординат $f(\xi_i)$ и еще пять весов w'_i производных $f'(\xi_i)$ ¹⁾:

$$\begin{array}{lll} \xi = -0,90, & w = 0,23640530, & w' = -0,00155377, \\ & -0,54, & 0,47899553, & 0,00058042, \\ & 0, & 0,56919830, & 0, \\ & 0,54, & 0,47899553, & -0,00058042, \\ & 0,90, & 0,23640530, & 0,00155377. \end{array}$$

В нашем примере изменение пределов приводит к тому, что точками интерполяции являются

$$x = 0,2; \quad 0,92; \quad 1; \quad 3,08; \quad 3,8;$$

последовательные ординаты и их производные будут

$$\begin{aligned} y_1 = y'_1 &= 1,22140276, \\ y_2 = y'_2 &= 2,50929030, \\ y_3 = y'_3 &= 7,38905610, \\ y_4 = y'_4 &= 21,75840240, \\ y_5 = y'_5 &= 44,70118449. \end{aligned}$$

Веса w_i ординат должны быть умножены на 2, а веса w'_i производных — на $2^2 = 4$. Умножив ординаты и их производные на соответствующие веса и просуммировав, получим в результате

$$\bar{A} = 53,37259512 + 0,22554004 = 53,59813516.$$

Погрешность теперь составляет

$$\tau_1 = 0,00001488.$$

Отсюда видно, что *точность* метода Гаусса сохранилась *полностью*.

Мы снова можем оценить погрешность с помощью метода, описанного в § 12. При этом опять получаем формулу (12.4), но со сле-

¹⁾ В настоящее время нет систематической таблицы весов w_i и w'_i для операций с ординатами и их производными при округленных гауссовых нулях.

дующим видоизменением: поправка в $\overline{A}$, вызванная весами w'_i , умножается на 2. Кроме того, следует добавить еще один член следующего вида:

$$- \left(\frac{b-a}{2} \right)^2 \sum_{\alpha=1}^n w'_\alpha \xi_\alpha f''(x_\alpha).$$

В нашем численном примере мы получаем

$$\eta' = 111,19630006 - 53,59813516 - 0,22554004 - \\ - 4 \cdot 14,229894611 - 8 \cdot 0,0566116792 = 0,000153.$$

Деление на $2n + 1 = 11$ дает $\eta = 0,0000139$. Согласие с фактической ошибкой опять удовлетворительное.

16. Применения квадратурного метода Гаусса в технике

Квадратурный метод Гаусса характеризуется весьма высокой степенью точности. Даже небольшое число ординат обыкновенно дает весьма точное значение определенного интеграла. В технических задачах редко требуется исключительная точность. Однако квадратурный метод Гаусса находит там применение как превосходный прием экономии числа ординат. Очень часто случается, что надо определить среднее значение функции неизвестной структуры на основе весьма немногих наблюдений. В этом случае точки, в которых измеряются ординаты, полезно выбирать согласно схеме Гаусса.

Например, в использованном нами численном примере нахождения определенного интеграла

$$A = \int_0^4 e^x dx$$

применение правила Симпсона при использовании девяти равноотстоящих ординат дало ошибку в 0,02 на 54 единицы, т. е. точность в 0,04%. Такая точность редко требуется в технических задачах.

Сократим теперь число ординат до *трех*. Схема Гаусса требует, чтобы эти ординаты были взяты при следующих значениях x и со следующими весами:

$$\begin{array}{ll} x = 0,45, & w = 1,11, \\ & 2, & 1,78, \\ & 3,55, & 1,11. \end{array}$$

Вычисление дает $\overline{A} = 53,535$ по сравнению с точным значением $A = 53,598$. Погрешность $\eta = 0,063$. Эта погрешность в три раза больше ошибки, получившейся при использовании девяти ординат, но все же приемлема, хотя взяты были только три ординаты. Следовательно, использование квадратурного метода Гаусса особенно

целесообразно, если по каким-либо соображениям надо экономить число употребляемых ординат для установления среднего значения неизвестной функции.

Если бы мы воспользовались в приведенном примере тремя *равноотстоящими* ординатами, то получили бы значение

$$\bar{A} = 56,77.$$

Теперь ошибка $\eta = 3,2$. Таким образом, погрешность в 50 раз больше, чем при пользовании методом Гаусса.

Напорные трубки в воздушной трубе дадут гораздо более полезные результаты, если они распределены по поперечному сечению трубы не равномерно, а в соответствии с узлами Гаусса. То же самое справедливо и для измерений температуры вдоль вала или измерений температуры на протяжении известного промежутка времени, если целью этих измерений является установление среднего значения.

Гауссовы узлы и соответствующие им веса вычислены с большой точностью и имеются в виде таблицы. Табл. XIII приложения, содержащая эти данные, рассчитана в Проектном бюро математических таблиц (New York). Там же были вычислены таблицы весов, соответствующих округленным значениям гауссовых нулей. Часть этой таблицы дана в приложении (см. табл. XIV).

17. Формула Симпсона с концевой поправкой

Мы видели, что погрешность, обусловленную небольшим смещением гауссовых узлов, можно было возместить добавлением к заданным значениям функции значений ее производных в тех же точках. Веса производных в симметрически расположенных точках входят со знаками $+$ и $-$. Этим свойством весов w'_k можно воспользоваться для такого видоизменения формулы Симпсона, которое намного увеличивает степень ее точности за счет небольшого количества добавочных вычислений. В § 4 мы разобрали метод Симпсона. Он сводится к тому, что мы подразделяем весь интервал на четное число полос и аппроксимируем каждую двойную полосу с помощью параболы второй степени. Сосредоточим внимание на одной такой двойной полосе и примем среднюю и две крайние ее точки за *двойные*. В результате этого мы получим аппроксимирование параболой *пятой степени*, и погрешность будет пропорциональна шестой производной, вместо четвертой производной в прежней формуле. Для достаточно гладких функций точность формулы, таким образом, намного возрастет. С другой стороны, теперь надо знать в каждой узловой точке значение функции и ее *производной*. Фактически, однако, так как каждая такая точка (за исключением двух крайних) является конечной точкой одной полосы и в то же время начальной точкой следующей полосы, производные входят с весами w' и $-w'$, сумма которых

равна нулю; только две концевые точки имеют разные веса. Следовательно, надо знать лишь производные в этих двух точках.

Веса вычисляются в этом случае столь просто, что нет нужды прибегать к помощи общего процесса § 13. Мы можем непосредственно решить линейные уравнения для весов, используя тот факт, что для степеней $x^0, x^1, \dots, x^5$ мы должны получить точные результаты. Нечетными степенями можно при этом пренебречь, так как для них уравнения автоматически выполняются в силу симметрии. Следовательно, использованию подлежат лишь значения функций x^0, x^2, x^4 . Таким образом, мы имеем 3 неизвестные, а именно два веса:

$$\frac{x = -1 \quad 0 \quad 1;}{w_1, w_0, w_1}$$

и третий вес

$$-w'_1, 0, w'_1.$$

Вес w'_0 можно заранее приравнять нулю в силу симметрии. Теперь мы имеем для трех функций x^0, x^2, x^4 :

$$\begin{aligned} y = x^0 & \begin{cases} y(-1) = 1, & y(0) = 1, & y(1) = 1, \\ y'(-1) = 0, & & y'(1) = 0, \end{cases} \\ y = x^2 & \begin{cases} y(-1) = 1, & y(0) = 0, & y(1) = 1, \\ y'(-1) = -2, & & y'(1) = 2, \end{cases} \\ y = x^4 & \begin{cases} y(-1) = 1, & y(0) = 0, & y(1) = 1, \\ y'(-1) = -4, & & y'(1) = 4. \end{cases} \end{aligned}$$

Это дает три уравнения

$$\begin{aligned} 2w_1 + w_0 &= 2, \\ 2w_1 + 4w'_1 &= \frac{2}{3}, \\ 2w_1 + 8w'_1 &= \frac{2}{5}, \end{aligned} \tag{6-17.1}$$

откуда

$$w'_1 = -\frac{1}{15}, \quad w_1 = \frac{7}{15}, \quad w_0 = \frac{16}{15}. \tag{6-17.2}$$

Мы получаем окончательную формулу:

$$\bar{A} = \frac{1}{15} [7(f(-1) + f(1)) + 16f(0) + f'(-1) - f'(1)]. \tag{6-17.3}$$

Если расстояние между соседними ординатами равно не 1, а h , то формулу надо заменить следующей:

$$\bar{A} = \frac{h}{15} [7(y_0 + y_2) + 16y_1 + h(y'_0 - y'_2)], \tag{6-17.4}$$

где y_0, y_1, y_2 обозначают три последовательные ординаты.

Пример. Функция

$$y = \sin x$$

между $x=0$ и π имеет, грубо говоря, вид параболы. Поэтому мы можем получить удовлетворительное приближение этой функции, задав ее в трех равноотстоящих точках

$$\begin{array}{c} x=0, \frac{\pi}{2}, \pi \\ \hline y=0, 1, 0 \end{array}$$

и аппроксимируя параболой второй степени. В этом случае

$$h = \frac{\pi}{2}, \quad y_0 = 0, \quad y_1 = 1, \quad y_2 = 0.$$

Кроме того,

$$y'_0 = 1, \quad y'_2 = -1.$$

Правило Симпсона дает

$$\bar{A} = \frac{\pi}{6} (y_0 + 4y_1 + y_2) = \frac{2\pi}{3} = 2,094,$$

тогда как точная величина площади

$$A = \int_0^{\pi} \sin x dx = -\cos x \Big|_0^{\pi} = 2.$$

Погрешность составляет, таким образом, $4,7\%$; $\eta = -0,094$. Новая формула дает для той же площади

$$\bar{A} = \frac{\pi}{30} (16 + \pi) = 2,0045.$$

Погрешность $\eta = -0,0045$, т. е. в 20 раз меньше, чем прежде.

Это увеличение точности можно объяснить, сопоставив оценки погрешности. Для формулы Симпсона эта оценка, согласно теореме интегрального исчисления о среднем, составляет

$$\eta = -\frac{f^{IV}(\theta)}{4!} \int_{-1}^{+1} x^2 (1-x^2) dx = -\frac{f^{IV}(\theta)}{90} \quad (6-17.5)$$

и, в общем случае (ширина полосы h , границы a и b),

$$\eta = -\frac{f^{IV}(\theta)}{180} h^4 (b-a). \quad (6-17.6)$$

Между тем, в случае формулы (3) мы получаем

$$\eta = \frac{f^{(6)}(\theta)}{6!} \int_{-1}^{+1} x^2 (1-x^2)^2 dx = \frac{f^{(6)}(\theta)}{4725}, \quad (6-17.7)$$

а для общего случая

$$\eta_1 = \frac{f^{(6)}(\theta)}{9450} h^6 (b - a). \quad (6-17.8)$$

Это дает, в приложении к нашему численному примеру, оценки:

$$\bar{\eta}_1 = -0,106 \text{ (обычная формула Симпсона),}$$

$$\bar{\eta}_1 = -0,00499 \text{ (видоизмененная формула Симпсона).}$$

Большая точность этих оценок погрешности объясняется тем фактом, что в нашем примере точка максимума производной $f^{(n)}(\theta)$ и середина интервала *совпадают*.

Разделим теперь заданный интервал на четное число полос, как это мы делали раньше при образовании параболического приближения Симпсона. Применим нашу формулу к каждой двойной полосе и сложим площади. Поправочные члены, содержащие y' , взаимно уничтожаются во всех внутренних точках, так как они входят в каждую смежную пару с противоположными знаками. Остаются только поправки, соответствующие двум *концевым точкам* интервала. Окончательная формула имеет вид

$$\bar{A} = \frac{h}{15} \left[14 \left(\frac{1}{2} y_0 + y_2 + y_4 + \dots + \frac{1}{2} y_{2n} \right) + 16 (y_1 + y_3 + y_5 + \dots + y_{2n-1}) + h (y'_0 - y'_{2n}) \right]. \quad (6-17.9)$$

Эта формула показывает, что сравнительно небольшой ценой (добавлением производных в двух конечных точках интервала) мы весьма значительно выигрываем в точности. Наше приближение теперь оказывается пятой степени. Это значит, что формула является точной для любой функции $f(x)$, которая может быть представлена каким-либо полиномом пятой степени. Правило Симпсона дает точные результаты лишь для полиномов третьей степени. В случае гладких функций две дополнительные степени создают большое увеличение точности.

Численный пример. Мы снова возвратимся к численному примеру § 8 и применим формулу (9) к девяти ординатам этой задачи. Теперь $h = \frac{1}{2}$ и $y' = e^x$. Если подставим в (9) численные значения, то получим

$$\bar{A} = 53,5980641$$

при точном значении

$$A = 53,5981500.$$

Ошибка, следовательно, равна $\eta = 0,000086$. Формула Симпсона давала ошибку $-0,0181$. Теперь погрешность в 200 раз меньше. Этот

пример наглядно иллюстрирует высокую эффективность концевой поправки. Оценка увеличения точности дается отношением

$$-\frac{f^{(6)}(\theta)}{f^{(4)}(\theta_1)} \frac{90h^2}{4725}.$$

Это отношение для взятого примера равно $\frac{1}{210}$, т. е. оценка погрешности скорректированной формулы в 210 раз меньше, чем оценка погрешности нескорректированной формулы, что хорошо согласуется с фактическим положением.

18. Квадратура, содержащая показательные функции

Во многих задачах прикладного анализа встречается «интегральное преобразование» следующего вида:

$$F(p) = \int_a^b f(x) e^{px} dx. \quad (6-18.1)$$

Мы принимаем, что функция $f(x)$ задается таблично, причем x изменяется через равные интервалы $\Delta x = h$. Мы принимаем также, что этот интервал настолько мал, что линейная интерполяция оказывается достаточной для вычисления значений функций в точках, лежащих между табличными точками. В этом случае мы могли бы воспользоваться простым правилом трапеций (3.5) для вычисления интеграла (1), если бы не показательный множитель e^{px} . Если бы p было достаточно мало, то правило трапеций все же оставалось бы пригодным. Но может понадобиться вычислить интеграл (1) при больших значениях p . Поэтому полезно знать, как вычислить интеграл вида (1) для функции $f(x)$, заданной таблично через достаточно малые интервалы без наложения каких-либо ограничений на p (эта величина может принимать любые вещественные, мнимые или комплексные значения).

Интерполируем линейно $f(x)$ последовательно в каждой полосе, начиная с ее средней точки $x_k + \frac{1}{2}h$ до двух точек $\left(x_k + \frac{1}{2}h\right) \pm \frac{1}{2}h$. Затем выполняем в каждой полосе интегрирование и берем сумму. В результате получаем формулу трапеций, но с некоторыми поправками. Пусть результатом процесса суммирования по правилу трапеций будет $S(p)$. Тогда

$$F(p) = \left[\sigma\left(\frac{p}{2}\right)\right]^2 S(p) + \frac{\sigma(p) - 1}{p} [f(a) e^{pa} - f(b) e^{pb}], \quad (6-18.2)$$

где

$$\sigma(p) = \frac{\sin ph}{ph}. \quad (6-18.3)$$

Эта формула имеет много приложений и особенно полезна при нахождении коэффициентов Фурье эмпирически заданной функции (в этом случае p — чисто мнимое число). Нескорректированная сумма $S(p)$ дает коэффициенты конечного тригонометрического ряда, который проходит через заданные точки (ср. IV, 11—15). Следовательно, формулу (2) можно понимать как выражение, устанавливающее связь между точными коэффициентами Фурье (которые требуются, например, в акустическом анализе эмпирически заданной функции) и коэффициентами, полученными при тригонометрической интерполяции.

19. Квадратура с помощью дифференцирования

Многие функции прикладного анализа определяются дифференциальным уравнением. Если нужно интегрировать такого рода функцию, то представляется желательным использовать для этого только значения функции и ее производных в двух концевых точках интервала, так как значения последовательных производных легко могут быть вычислены из определяющего функцию дифференциального уравнения, если мы знаем граничные значения в двух концевых точках. Наша задача поэтому состоит в том, чтобы получить эффективную квадратурную формулу, которая пользуется не внутренними ординатами, а только двумя краевыми ординатами и значениями производных в этих точках. Такого рода формула использует краевую информацию не для увеличения точности, а для вычисления определенного интеграла. Можно также сказать, что наши ординаты теперь распределяются особым образом, поскольку они скапливаются бесконечно близко к обоим краевым точкам заданного интервала. Обычное разложение Тейлора соответствует случаю, когда все заданные ординаты лежат бесконечно близко к *одной* точке интервала. Мы же теперь принимаем, что *обе* конечные точки одинаково представлены.

Мы будем исходить из элементарной формулы интегрального исчисления, основанной на методе интегрирования по частям:

$$\begin{aligned} \int_a^b u(x) v^{(n)}(x) dx &= \int_a^b (-1)^n v(x) u^{(n)}(x) dx = \\ &= (uv^{(n-1)} - u'v^{(n-2)} + \dots) \Big|_a^b = \\ &= \left[\sum_{k=0}^{n-1} u^{(k)}(x) v^{(n-k-1)}(x) (-1)^k \right] \Big|_a^b. \quad (6-19.1) \end{aligned}$$

Этой формулой мы воспользуемся следующим образом. Промежуток интегрирования нормируем до промежутка $(0,1)$. Кроме того, мы полагаем

$$u = f(x), \quad v = \frac{p_n(x)}{\gamma_n^n n!}, \quad (6-19.2)$$

причем полиномом

$$p_n(x) = \gamma_n^n x^n + \gamma_{n-1}^n x^{n-1} + \dots + \gamma_0^n \quad (6-19.3)$$

мы можем свободно распорядиться.

При этом выборе функций $u(x)$ и $v(x)$ формула (1) может быть записана следующим образом:

$$\int_0^1 f(x) dx = \frac{1}{\gamma_n^n n!} \left[\sum_{k=0}^{n-1} f^{(k)}(x) p_n^{(n-k-1)}(x) (-1)^k \right]_0^1 + \eta_n, \quad (6-19.4)$$

где η_n означает определенный интеграл:

$$\eta_n = (-1)^n \int_0^1 \frac{p_n(x)}{\gamma_n^n n!} f^{(n)}(x) dx. \quad (6-19.5)$$

Формулу (4) следует понимать как *квадратурную формулу*, которая выражает площадь под кривой исключительно через краевые значения и производные, заданные в двух конечных точках $x=0$ и $x=1$ интервала:

$$\bar{A}_n = \frac{1}{\gamma_n^n n!} \left[\sum_{k=0}^{n-1} f^{(k)}(x) p_n^{(n-k-1)}(x) (-1)^k \right]_0^1, \quad (6-19.6)$$

в то время как величина η_n , заданная в форме (5), представляет *остаток* квадратурной формулы.

Обратимся сначала к этому остатку. Нашей целью будет распорядиться полиномом $p_n(x)$ таким образом, чтобы сделать остаток (5) возможно малым. Один способ подбора полинома $p_n(x)$ представляет особый интерес для данного случая, так как он переносит характерные черты квадратурного метода Гаусса в нашу задачу. При рассмотрении метода Гаусса мы видели (ср. § 10), что произвольное распределение точек интерполяции приводит к квадратурной формуле, которая дает точные результаты для произвольного полинома не выше $(n-1)$ -й степени, тогда как гауссовы точки интерполяции привели к квадратурной формуле, которая дает точные результаты для любого полинома $(2n-1)$ -й степени.

В рассматриваемом случае мы не можем выбирать точки интерполяции, ибо мы уже решили, что наша квадратура будет базироваться на граничных значениях заданной функции и ее производных на обоих концах интервала. Но полиномом $p_n(x)$ мы можем еще свободно распорядиться. Выражение (5) для остатка показывает, что наша квадратурная формула будет точной для всякого полинома не выше $(n-1)$ -й степени, но остаток не будет, вообще говоря, равняться нулю для полинома более высокой степени. При одном частном выборе полинома $p_n(x)$, однако, мы можем сделать

квадратурную формулу (4) точной для любого полинома вплоть до $(2n - 1)$ -й степени включительно.

Рассмотрим полиномы Лежандра $P_n(x)$ (см. V, 20), приведенные, однако, к интервалу $(0, 1)$ вместо традиционного интервала $(-1, 1)$. Эти полиномы мы обозначим через $P_n^*(x)$. Они могут быть непосредственно выражены через гауссову гипергеометрическую функцию $F(\alpha, \beta, \gamma, x)$ [см. (5-20.11)]:

$$\begin{aligned} P_n^*(x) &= F(-n, n+1, 1; x) = \\ &= 1 - \frac{n(n+1)}{1 \cdot 1} \cdot x + \frac{n(n-1)(n+1)(n+2)}{1 \cdot 2 \cdot 1 \cdot 2} x^2 - \dots = \\ &= \sum_{k=0}^n \frac{(n+k)!}{(n-k)!(k!)^2} (-x)^k. \end{aligned} \quad (6-19.7)$$

Полином $P_n^*(x)$ обладает следующим замечательным свойством. Он может быть записан в виде n -й производной от функции, которая обращается в нуль вместе со всеми своими производными вплоть до $(n-1)$ -го порядка в обеих конечных точках интервала $x=0$ и $x=1$:

$$P_n^*(x) = \frac{1}{n!} \frac{d^n [x(1-x)]^n}{dx^n}. \quad (6-19.8)$$

Но тогда мы можем воспользоваться интегральным преобразованием (1), отождествив $u(x)$ с $f^{(n)}(x)$ и $v(x)$ с $[x(1-x)]^n$. Краевые члены в правой части полностью выпадают в силу свойств функции $v(x)$, а остаток может быть записан в виде

$$\begin{aligned} \eta_n &= \frac{(-1)^n}{\gamma_n^n (n!)^2} \int_0^1 f^{(n)}(x) \frac{d^n [x(1-x)]^n}{dx^n} dx = \\ &= \frac{1}{\gamma_n^n (n!)^2} \int_0^1 f^{(2n)}(x) [x(1-x)]^n dx. \end{aligned} \quad (6-19.9)$$

По определению, γ_n^n обозначает коэффициент при старшем члене полинома $P_n^*(x)$. Согласно (7) мы поэтому получаем

$$\gamma_n^n = (-1)^n \frac{(2n)!}{(n!)^2}. \quad (6-19.10)$$

Кроме того, полином Лежандра $P_n^*(x)$ обладает свойством симметрии:

$$P_n^*(x) = (-1)^n P_n^*(1-x). \quad (6-19.11)$$

В силу этого свойства сумма в правой части формулы (6) может

быть сведена с $2n$ до n членов:

$$\bar{A}_n = \frac{1}{\gamma_n^n n!} \sum_{k=0}^{n-1} (-1)^{k+1} P_n^{*(n-k-1)}(0) y_k, \quad (6-19.12)$$

где

$$y_k = f^{(k)}(0) + (-1)^k f^{(k)}(1). \quad (6-19.13)$$

Но разложение (7) дает

$$P_n^{*(k)}(0) = (-1)^k \frac{(n+k)!}{(n-k)! k!}. \quad (6-19.14)$$

Введем обозначение

$$(-1)^{n-k} P_n^{*(n-k)}(0) = D_k^n = \frac{(2n-k)!}{(n-k)! k!}. \quad (6-19.15)$$

Тогда окончательная квадратурная формула принимает вид

$$\bar{A}_n = \frac{1}{D_0^n} \sum_{k=0}^{n-1} D_{k+1}^n y_k. \quad (6-19.16)$$

Коэффициенты D_k^n приведены в табл. XV приложения.

Для произвольного промежутка $[a, b]$ формула (16) видоизменяется следующим образом. Мы вводим обозначение

$$y_k = f^{(k)}(a) + (-1)^k f^{(k)}(b) \quad (6-19.17)$$

и получаем

$$\bar{A}_n = \frac{1}{D_0^n} \sum_{k=0}^{n-1} D_{k+1}^n (b-a)^{k+1} y_k. \quad (6-19.18)$$

В простейшем частном случае, когда функция и ее первая и вторая производные заданы в обеих концевых точках интервала, причем $n=3$, мы получаем

$$\begin{aligned} \bar{A} = \frac{1}{2} f(a) h + \frac{1}{10} f'(a) h^2 + \frac{1}{120} f''(a) h^3 + \frac{1}{2} f(b) h - \\ - \frac{1}{10} f'(b) h^2 + \frac{1}{120} f''(b) h^3 \end{aligned}$$

при $h = b - a$.

Погрешность η_n квадратурной формулы (16) можно оценить следующим образом. Так как весовая функция $[x(1-x)]^n$ не меняет знака на всем интервале $[0, 1]$, то по теореме о среднем взвешенном имеем

$$\frac{\int_0^1 f^{(2n)}(x) [x(1-x)]^n dx}{\int_0^1 [x(1-x)]^n dx} = f^{(2n)}(\theta), \quad (6-19.19)$$

где θ есть некоторая точка в интервале $[0, 1]$. В силу того, что

$$\int_0^1 [x(1-x)]^n dx = \frac{\sqrt{\pi}}{2^{2n+1}} \frac{n!}{\left(n + \frac{1}{2}\right)!}, \quad (6-19.20)$$

имеем

$$\bar{r}_n = \frac{(-1)^n \sqrt{\pi}}{2 \cdot 4^n} \cdot \frac{n!}{\left(n + \frac{1}{2}\right)! (2n)!} f^{(2n)}(\theta). \quad (6-19.21)$$

Для случая произвольных пределов

$$r_n = \sqrt{\pi} \frac{(-1)^n n!}{\left(n + \frac{1}{2}\right)! (2n)!} f^{(2n)}(\theta) \left(\frac{b-a}{2}\right)^{2n+1}. \quad (6-19.22)$$

Мы можем избавиться от неопределенности положения точки θ , если имеется возможность принять, что $f^{(2n)}(x)$ не изменяется слишком сильно внутри интервала $[a, b]$. Функция $[x(1-x)]^n$ вырезает узкое «окно» вокруг точки $x = \frac{1}{2}$, ибо при достаточно большом n эта функция быстро убывает по обе стороны от этой точки. Поэтому среднее взвешенное значений (19) сосредоточено главным образом в непосредственной окрестности точки $x = \frac{1}{2}$. Если при этом $f^{(2n)}(x)$ достаточно гладко изменяется, то мы можем отождествить θ с точкой $x = \frac{1}{2}$, а в общем случае — с точкой

$$\theta = \frac{b-a}{2}.$$

Кроме того, формула Стирлинга для факториала показывает, что при оценках можно положить

$$\frac{n!}{\left(n + \frac{1}{2}\right)!} = n^{-1/2}. \quad (6-19.23)$$

При этих условиях мы получаем следующую реалистическую оценку погрешности квадратурной формулы (18):

$$\bar{\eta}_n = (-1)^n \sqrt{\frac{\pi}{n}} \frac{f^{(2n)}\left(\frac{b-a}{2}\right)}{(2n)!} \left(\frac{b-a}{2}\right)^{2n+1}. \quad (6-19.24)$$

Эта формула весьма сходна с формулой, относящейся к квадратуре Гаусса¹⁾ [ср. (12.1)]. Однако численные множители значительно

¹⁾ Поэтому метод оценки погрешности, изложенный в § 12, применим и здесь. В качестве вспомогательной функции берем $[(2x-1)f(x)]'$ (если взят интервал $[0, 1]$), для которой результатом квадратуры является $f(1) + f(0)$. Следовательно, мы снова располагаем явным выражением погрешности η' .

различаются. Отношение этих множителей равно

$$\mu_n = \frac{(-4)^n}{V \pi n}. \quad (6-19.25)$$

Таким образом, судя по оценкам, погрешность квадратурной формулы, основанной на $2n$ краевых значениях, в μ_n раз больше соответствующей погрешности квадратуры Гаусса, использующей n внутренних точек.

Это обстоятельство вполне объяснимо тем, что выбор информации исключительно в двух концах интервала значительно менее выгоден, чем разумный выбор ординат на всем интервале. Однако важное значение этого метода состоит в том, что часто гораздо легче определить функцию и ее производные в двух концевых точках интервала, чем вычислить значения функции в нескольких внутренних точках. Это, в частности, относится к случаю, когда неизвестная функция, для которой нет таблиц, определена *дифференциальным уравнением*. Поэтому применение квадратурной формулы (16) может дать эффективный метод *решения* заданного дифференциального уравнения (ср. § 21).

20. Примеры

Частным примером, хорошо иллюстрирующим силу рассмотренного в § 19 метода, может служить показательная функция

$$f(x) = e^{\alpha x}, \quad (6-20.1)$$

интегрируемая от 0 до 1. Здесь нам известен результат интегрирования

$$\int_0^1 e^{\alpha x} dx = \frac{1}{\alpha} (e^\alpha - 1). \quad (6-20.2)$$

Собрав все члены с e^α и заменив α через x , мы получим рациональное приближение показательной функции e^x в следующей форме¹⁾:

$$e^x = \frac{\sum_{k=0}^n D_k^n x^k}{\sum_{k=0}^n D_k^n (-x)^k}. \quad (6-20.3)$$

а оценка погрешности первоначальной квадратуры опять имеет вид:

$$\bar{\eta} = \frac{\eta'}{2n+1}.$$

¹⁾ Автор признателен своему другу Ч. Девису (C. Davis), обратившему внимание на то, что это приближение было уже найдено раньше (см. P. M. Hummel и C. L. Seebeck, A Generalization of Taylor's Theorem, Association Monthly, 56, 243—247, 1949).

Первые четыре приближения ($n = 1, 2, 3, 4$) выражаются так:

$$\begin{aligned} \frac{2+x}{2-x}; \quad \frac{12+6x+x^2}{12-6x+x^2}; \quad \frac{120+60x+12x^2+x^3}{120-60x+12x^2-x^3}; \\ \frac{1680+840x+180x^2+20x^3+x^4}{1680-840x+180x^2-20x^3+x^4}. \end{aligned} \quad (6-20.4)$$

Приближение имеет касание с кривой e^x порядка $2n$; это значит, что при разложении этих дробей по степеням x совпадение с коэффициентами разложения Тейлора для функции e^x имеет место вплоть до члена $2n$ -й степени, хотя в нашем распоряжении имеется лишь n коэффициентов.

Если положить $x=1$, то получаются последовательные рациональные приближения трансцендентного числа e , характеризующиеся удивительной точностью:

$$\begin{aligned} e &= \frac{3}{1} = 3, \\ \frac{19}{7} &= 2,714 \quad (\eta = 4 \cdot 10^{-3}), \\ \frac{193}{71} &= 2,71831 \quad (\eta = -3 \cdot 10^{-5}), \quad (6-20.5) \\ \frac{2721}{1001} &= 2,7182817 \quad (\eta = 1 \cdot 10^{-7}), \\ \frac{49171}{18089} &= 2,7182818287 \quad (\eta = -3 \cdot 10^{-10}), \\ \frac{1084483}{398959} &= 2,7182818284586 \quad (\eta = 4 \cdot 10^{-13}). \end{aligned}$$

Полезно также применить нашу квадратурную формулу к тому же численному примеру, которым мы воспользовались для иллюстрации эффективности квадратуры Гаусса (ср. § 11). Мы имели там $n=5$, $a=0$, $b=4$, а заданная функция была $y=e^x$. Применяя формулу (19.18) с коэффициентами D_k^5 , взятыми из табл. XV приложения, получим

$$\begin{aligned} \bar{A} = \frac{1}{30 \cdot 240} [15 \cdot 120 \cdot 4 (1 + e^4) + 3360 \cdot 16 (1 - e^4) + 420 \cdot 64 (1 + e^4) + \\ + 30 \cdot 256 (1 - e^4) + 1 \cdot 1024 (1 + e^4)] = 53,60174. \end{aligned}$$

Сравнение с точным значением $A=53,59815$ показывает, что

$$\eta = -0,0036,$$

тогда как погрешность гауссовой формулы составляла всего лишь $\eta=0,000013$. Изменение знака объясняется множителем $(-1)^n$ формулы (19.24), который в случае $n=5$ дает знак минус. Далее,

множитель μ_n , являющийся оценкой увеличения погрешности по сравнению с погрешностью квадратуры Гаусса, теперь равен

$$|\mu_5| = \frac{4^5}{\sqrt{5\pi}} = 258,$$

что хорошо согласуется с фактическим увеличением.

21. Задачи собственных значений

В §§ 17 и 18 гл. V мы встретились с классом задач, играющих фундаментальную и все возрастающую роль во всякого рода проблемах вибрации, связанных с упругостью, анализом неустановившихся колебаний, волноводами и атомной физикой. Это — так называемые «задачи собственных значений». Общее положение, встречающееся в таких задачах, можно охарактеризовать следующим образом. Дано некоторое линейное дифференциальное уравнение, содержащее неизвестный постоянный параметр λ , обычно называемый «собственным значением»; даны также некоторые однородные краевые условия такого рода, что решение уравнения (если не считать тривиального решения $y = 0$) возможно лишь при надлежащем выборе параметра λ ; задача состоит в том, чтобы найти наименьшее значение λ_1 или несколько наименьших значений λ_i , которые делают возможным решение уравнения.

В таких задачах на собственные значения квадратурный метод § 19 может оказать существенную помощь, так как он базируется на знании функции и ее производных в двух краевых точках интервала, а эти величины могут быть получены на основе заданного дифференциального уравнения и заданных краевых условий.

Для того чтобы проиллюстрировать использование этого метода, мы выберем простой пример; но метод применим при гораздо более сложных условиях. Наша цель, однако, изучить существенные черты метода, не осложненные техническими трудностями. Поэтому мы остановимся на простом дифференциальном уравнении второго порядка с постоянными коэффициентами:

$$y'' + y' + \lambda y = 0 \tag{6-21.1}$$

с краевыми условиями

$$y(0) = 0, \quad y'(1) = 0. \tag{6-21.2}$$

Заданный интервал, таким образом, приведен к $[0, 1]$.

Так как линейное однородное дифференциальное уравнение оставляет амплитудный множитель неопределенным, то можно приписать производной в точке $x = 0$ произвольное значение

$$y'(0) = 1.$$

При этих условиях дифференциальное уравнение (1) однозначно определяет значения всех производных в точке $x=0$. Мы получаем их последовательным дифференцированием, либо подстановкой в дифференциальное уравнение разложения y в степенной ряд

$$y = a_0 + a_1 x + a_2 x^2 + \dots$$

Собираем подобные члены и приравниваем полученные коэффициенты при x^k нулю. В нашем простом примере находим

$$\begin{aligned} a_0 = 0, \quad a_1 = 1, \quad a_2 = -\frac{1}{2}, \quad a_3 = \frac{1-\lambda}{6}, \\ y(0) = 0, \quad y'(0) = 1, \quad y''(0) = -1, \quad y'''(0) = 1 - \lambda. \end{aligned} \quad (6-21.3)$$

Чем дальше мы продолжим вычисление последовательных коэффициентов, тем большей точности можем ожидать. Для наших целей мы остановимся здесь на a_3 .

В другой конечной точке мы полагаем аналогично

$$x = 1 + \xi, \quad y = b_0 + b_1 \xi + b_2 \xi^2 + b_3 \xi^3.$$

Пользуясь тем же методом подстановки в дифференциальное уравнение и приведения членов, мы не получим значения b_0 , но все последующие коэффициенты оказываются линейными функциями коэффициента b_0 :

$$\begin{aligned} b_0, \quad b_1 = 0, \quad b_2 = -\frac{\lambda b_0}{2}, \quad b_3 = \frac{\lambda b_0}{6}, \\ y(1) = b_0, \quad y'(1) = 0, \quad y''(1) = -\lambda b_0, \quad y'''(1) = \lambda b_0. \end{aligned} \quad (6-21.4)$$

Теперь мы применим квадратурную формулу, рассматривая $y''(x)$ как начальную функцию $f(x)$; следовательно, $f(x) = y''(x)$ и $f'(x) = y'''(x)$ заданы в обеих конечных точках:

$$\begin{aligned} f(0) = -1, \quad f(1) = -\lambda b_0, \\ f'(0) = 1 - \lambda, \quad f'(1) = \lambda b_0. \end{aligned}$$

Квадратурная формула при $n=2$ дает

$$\begin{aligned} \int_0^1 y''(x) dx = y'(1) - y'(0) = \frac{6(-1 - \lambda b_0) + 1(1 - \lambda - \lambda b_0)}{12}; \\ -1 = \frac{-5 - 7\lambda b_0 - \lambda}{12}. \end{aligned}$$

Это приводит к соотношению

$$b_0 = \frac{7 - \lambda}{7\lambda}. \quad (6-21.5)$$

Теперь мы еще раз воспользуемся квадратурной формулой, рассматривая $y'(x)$ в качестве $f(x)$:

$$\begin{aligned} f(0) &= 1, & f(1) &= 0, \\ f'(0) &= -1, & f'(1) &= -\lambda b_0, \\ f''(0) &= 1 - \lambda, & f''(1) &= \lambda b_0. \end{aligned}$$

Сейчас мы имеем три пары данных в обоих концах и поэтому применяем формулу при $n=3$:

$$\int_0^1 y'(x) dx = y(1) - y(0) = \frac{60(1+0) + 12(-1 + \lambda b_0) + (1 - \lambda + \lambda b_0)}{120},$$

$$b_0 = \frac{49 - \lambda + 13\lambda b_0}{120}.$$

Это дает новое соотношение:

$$b_0 = \frac{49 - \lambda}{120 - 13\lambda}. \quad (6-21.6)$$

Приравняв правые части равенств (5) и (6), мы получаем квадратное уравнение для определения собственного значения λ :

$$20\lambda^2 - 554\lambda + 840 = 0.$$

Два корня этого уравнения равны

$$\lambda_1 = 1,6095, \quad \lambda_2 = 26,0905. \quad (6-21.7)$$

Имеет значение лишь меньший корень, так как больший корень чрезвычайно сильно изменяется при введении в расчеты следующих членов разложения. Напротив, меньший корень не очень изменяется, если продолжить наш процесс дифференцирования. Мы остановились на $y'''(x)$. Если мы сделаем еще один шаг и включим в наши вычисления $y'''(0)$ и $y'''(1)$, то первое приближение квадратурной формулы будет при $n=3$, а второе — при $n=4$. Мы получаем при этом два соотношения:

$$b_0 = \frac{71 - 10\lambda}{\lambda(73 - \lambda)} \quad \text{и} \quad b_0 = \frac{679 - 18\lambda}{1680 - 201\lambda + \lambda^2},$$

которые дают теперь для определения λ следующее кубическое уравнение:

$$28\lambda^3 - 4074\lambda^2 + 80\,638\lambda - 119\,280 = 0.$$

Это все-таки по существу уравнение квадратное, так как кубический член представляет собой лишь небольшую поправку. Мы можем сначала пренебречь этим кубическим членом и получить предварительное значение λ_0 , а затем добавить $28\lambda_0^3$ в качестве поправки к свободному члену. Это дает для наименьшего корня значение

$$\lambda_1 = 1,608467. \quad (6-21.8)$$

Малое изменение величины λ_1 по сравнению с соответствующим значением, найденным в (7), показывает, что первое довольно грубое приближение было очень близким к действительности, так как ошибка составляла лишь одну единицу третьего десятичного разряда, что представляет точность до $0,07\%$. В данном простом примере мы можем проверить наши результаты, так как можно теоретически получить точное значение λ_1 . Оно определяется формулой

$$\lambda = \frac{1}{4} + \theta^2, \quad (6-21.9)$$

где θ есть решение трансцендентного уравнения

$$\operatorname{tg} \theta = 2\theta. \quad (6-21.10)$$

Наименьший корень этого уравнения

$$\theta_1 = 1,1655618,$$

что дает

$$\lambda_1 = 1,608534. \quad (6-21.11)$$

Таким образом, погрешность приближения при $n=2,3$ составляет $\eta = -0,0010$, тогда как при $n=3,4$ погрешность равна лишь $\eta = 0,000067$. Следовательно, сходимость метода при применении его к гладкой функции — быстрая. Представляет интерес также приближение сверху и снизу в двух последовательных случаях.

Так называемый «метод Релея — Ритца» получения собственных значений основан на минимизировании некоторого интеграла. Поэтому он применим лишь к *самосопряженным* дифференциальным операторам. Настоящий же метод не требует, чтобы дифференциальный оператор либо краевые условия были самосопряженными. Он применим при более общих условиях, и, принципиально, даже линейный характер дифференциального уравнения необязателен для приложения этого метода.

Если следовать этой идее несколько дальше, мы придем к заключению, что рассмотренный квадратурный процесс может быть применен не только к вычислению собственных значений, но также и к действительному *решению* дифференциальных уравнений. Сделаем подстановку

$$x = \alpha x_1, \quad (6-21.12)$$

рассматривая x_1 как новую независимую переменную, а α как заданный постоянный параметр. Значение $x_1 = 1$ новой переменной соответствует значению $x = \alpha$ старой переменной. Следовательно, $y(1)$ в новой переменной дает фактически первоначальную функцию $y(\alpha)$, а заменив α через x , мы получим неизвестную функцию в переменной точке x .

Покажем, как используется этот метод на прежнем примере. Дифференциальное уравнение (1) теперь принимает вид

$$y'' + \alpha y' + \alpha^2 \lambda y = 0.$$

Таблица (3) начальных значений должна быть заменена следующей:

$$\begin{aligned} a_0 &= 0, & a_1 &= \alpha, & a_2 &= -\frac{1}{2} \alpha^2, & a_3 &= \frac{1-\lambda}{6} \alpha^3, \\ y(0) &= 0, & y'(0) &= \alpha, & y''(0) &= -\alpha^2, & y'''(0) &= (1-\lambda) \alpha^3, \end{aligned}$$

тогда как таблица конечных значений (4) изменяется следующим образом:

$$\begin{aligned} y(1) &= b_0, & y'(1) &= b_1, & y''(1) &= -\alpha b_1 - \alpha^2 \lambda b_0, \\ y'''(1) &= (1-\lambda) \alpha^2 b_1 + \alpha^3 \lambda b_0. \end{aligned}$$

В нашей задаче неизвестными являются b_0 и b_1 , тогда как λ уже известно. Однако процедура остается той же самой. Первое соотношение, которое раньше привело к равенству (5), теперь принимает вид

$$b_1 - \alpha = \frac{1}{12} [6(-\alpha^2 - \alpha b_1 - \alpha^2 \lambda b_0) + (1-\lambda) \alpha^3 - (1-\lambda) \alpha^2 b_1 - \alpha^3 \lambda b_0].$$

Мы приходим к линейному соотношению:

$$(6\alpha^2 \lambda + \alpha^3 \lambda) b_0 + [12 + 6\alpha + (1-\lambda) \alpha^2] b_1 = 12\alpha - 6\alpha^2 + (1-\lambda) \alpha^3.$$

Для второго соотношения мы упростим расчеты, жертвуя точностью: снова воспользуемся квадратурной формулой для $n=2$, применив ее к первым трем краевым значениям и не пользуясь значениями $y'''(0)$ и $y'''(1)$. Тогда получим

$$b_0 = \frac{1}{12} [6(\alpha + b_1) + (-\alpha^2 + \alpha b_1 + \alpha^2 \lambda b_0)],$$

что дает новое линейное соотношение

$$(12 - \alpha^2 \lambda) b_0 - (6 + \alpha) b_1 = 6\alpha - \alpha^2.$$

Решив эти два линейных уравнения относительно b_0 и b_1 , получим

$$\begin{aligned} b_0 &= \frac{\alpha(144 - 12\lambda\alpha^2)}{144 + 72\alpha + 12(1+\lambda)\alpha^2 + 6\lambda\alpha^3 + \lambda^2\alpha^4}, \\ b_1 &= \alpha \frac{144 - 72\alpha + 12(1-5\lambda)\alpha^2 + 6\lambda\alpha^3 + \lambda^2\alpha^4}{144 + 72\alpha + 12(1+\lambda)\alpha^2 + 6\lambda\alpha^3 + \lambda^2\alpha^4}. \end{aligned}$$

Теперь мы перейдем обратно к первоначальной переменной x ; в этой переменной значения b_0 и b_1 будут $b_0 = y(\alpha)$ и $b_1 = \alpha y'(\alpha)$. Таким образом, мы приходим к двум независимым приближениям для $y(x)$ и $y'(x)$ в форме:

$$\begin{aligned} \bar{y}(x) &= \frac{x(144 - 12\lambda x^2)}{144 + 72x + 12(1+\lambda)x^2 + 6\lambda x^3 + \lambda^2 x^4}, \\ \bar{y}'(x) &= \frac{144 - 72x + 12(1-5\lambda)x^2 + 6\lambda x^3 + \lambda^2 x^4}{144 + 72x + 12(1+\lambda)x^2 + 6\lambda x^3 + \lambda^2 x^4}. \end{aligned}$$

Вторая производная $y''(x)$ и все высшие производные тогда определяются из исходного дифференциального уравнения (1).

Чтобы испытать точность нашего решения, мы применим его к частному значению $\lambda = -2$; в этом случае точное решение нашей задачи будет иметь вид

$$y(x) = \frac{1}{3}(e^x - e^{-2x}), \quad y'(x) = \frac{1}{3}(e^x + 2e^{-2x}),$$

тогда как

$$\bar{y}(x) = \frac{x(144 + 24x^2)}{144 + 72x - 12x^2 - 12x^3 + 4x^4},$$

$$\bar{y}'(x) = \frac{144 - 72x + 132x^2 - 12x^3 + 4x^4}{144 + 72x - 12x^2 - 12x^3 + 4x^4}.$$

В точке $x = 1$ точное решение дает

$$y(1) = 0,8610, \quad y'(1) = 0,9963,$$

тогда как из приближенного находим

$$\bar{y}(1) = 0,8571, \quad \bar{y}'(1) = 1.$$

Как видим, точность весьма удовлетворительная.

Таким образом, применение этого квадратурного метода к аналитическому решению задачи с краевыми условиями в области обыкновенных дифференциальных уравнений можно описать следующим образом. Сначала мы надлежащим масштабным преобразованием приводим заданный промежуток к $[0,1]$. Затем составляем таблицу краевых значений и производных:

$$\begin{array}{l} y(0), y'(0), y''(0), \dots; \\ y(1), y'(1), y''(1), \dots; \end{array}$$

таблицу обрываем на производных такого порядка, что они уже определяются заданным дифференциальным уравнением. Некоторые из этих краевых значений находим из краевых условий задачи. Остальные обозначаем буквами. Если дифференциальное уравнение однородно, то одно краевое значение можно нормировать к 1. Затем применяем квадратурную формулу (19.16); эта формула приводит нас к установлению линейного соотношения между неизвестными краевыми значениями. Используя квадратурную формулу для производных все более низкого порядка, мы в конце концов определяем все краевые значения из системы линейных уравнений. Если в уравнения входит собственное значение λ , то исключение неизвестных краевых значений приводит к алгебраическому уравнению относительно λ ; наименьший корень этого уравнения мы находим.

Теперь мы имеем задачу с начальным значением, так как $y(0)$ и все высшие производные функции $y(x)$ в точке $x = 0$ уже известны. Теперь следует выполнить преобразование $x = \alpha \xi$, где α

рассматривается как заданный постоянный параметр. Применяя по-прежнему квадратурную формулу к новому дифференциальному уравнению, мы, наконец, получаем $y(\alpha)$ и независимо от этого $y'(\alpha)$, $y''(\alpha)$, ... из заданных начальных значений.

Ценным контролем может служить выполнение того же процесса для *другой конечной точки* $x=1$ и сравнение значения $y(\alpha)$, найденного с помощью конечных значений $y(1)$, $y'(1)$, ..., с ранее полученным $y(\alpha)$, выраженным через *начальные значения* $y(0)$, $y'(0)$, ... Степень совпадения дает хорошую оценку точности решения.

22. Сходимость квадратуры, использующей краевые значения

Перейдем теперь к исследованию общих свойств сходимости рассмотренного квадратурного процесса. Мы воспользуемся для этого методом, весьма сходным с тем, которым пользовались ранее, когда исследовали сходимость равноотстоящей полиномиальной интерполяции (ср. V, 15). Будем предполагать, что функция $f(x)$ имеет аналитический характер в интервале $[0, 1]$ и в некоторой области комплексной плоскости, охватывающей этот интервал. Тогда можно представить $f(x)$ с помощью контурного интеграла Коши (5-15.9). Это дает возможность ограничить все исследования только специальной функцией $(z_0 - x)^{-1}$, где z_0 есть некоторая фиксированная точка комплексной плоскости. Если мы сможем показать, что наш квадратурный метод сходится для этой частной функции при условии, что z_0 лежит вне некоторой вполне определенной области комплексной плоскости, то сходимость обеспечена для любой аналитической функции $f(z)$, которая остается регулярной внутри и на границе этой области.

В (19.9) мы вывели следующую форму остатка:

$$\eta_n = \frac{(-1)^n}{(2n)!} \int_0^1 f^{(2n)}(x) [x(1-x)]^n dx. \quad (6-22.1)$$

Для частной функции

$$f(x) = \frac{1}{z_0 - x} \quad (6-22.2)$$

получаем

$$\frac{f^{(2n)}(x)}{(2n)!} = \frac{1}{(z_0 - x)^{2n+1}} \quad (6-22.3)$$

и, следовательно,

$$\eta_n = (-1)^n \int_0^1 \frac{1}{z_0 - x} \left[\frac{x(1-x)}{(z_0 - x)^2} \right]^n dx. \quad (6-22.4)$$

Мы должны теперь доказать, что с возрастанием n до бесконечности η_n стремится к нулю. Заметим, что решающей величиной

является абсолютное значение дроби, n -я степень которой входит в подынтегральное выражение:

$$A(x) = \left| \frac{x(1-x)}{(z_0-x)^2} \right|. \quad (6-22.5)$$

Если $A(x)$ повсюду остается меньше единицы, то постепенное уменьшение η_n до нуля обеспечено.

Может случиться, что точка z_0 настолько удалена от интервала $[0, 1]$ оси x , что $A(x)$ на всем интервале остается меньше, чем 1; в этом случае сходимость уже гарантирована и никакого дальнейшего исследования не требуется. Возможно, однако, что точка z_0 настолько близка к оси x , что $A(x)$ оказывается больше 1 в некоторой части нашего интервала. Допустим, что z_0 имеет вид

$$z_0 = \alpha - i\beta, \quad (6-22.6)$$

где β положительно, т. е. что z_0 есть некоторая комплексная точка, расположенная ниже оси x . Начиная с точки $x=0$, в которой $A(x)$ равно нулю, перемещаемся по оси x до точки P_1 , где $A(x)$ становится равной 1. Подобным же образом мы, начиная с точки $x=1$ (где $A(x)$ также равно нулю), возвращаемся назад до точки P_2 , в которой $A(x)$ опять равно 1. Затруднение представляет лишь часть интервала между точками P_1 и P_2 , так как части интервала, соответствующие отрезкам интервала $0P_1$ и P_21 , стремятся к нулю. Но теперь мы воспользуемся известным свойством аналитических функций, заключающимся в том, что путь интегрирования можно произвольно деформировать, если только не захватывать при этом особых точек подынтегральной функции. Поэтому мы заменяем x комплексной переменной $z = x + iy$ и выбираем путь интегрирования где-нибудь в верхней половине комплексной плоскости, где подынтегральная функция не имеет особых точек. Сначала определим геометрическое место точек, в которых $A(x)$ равно 1. Это приводит к условию

$$[x(1-x) + y^2]^2 + (1-2x)^2 y^2 = [(\alpha-x)^2 + (\beta+y)^2]^2, \quad (6-22.7)$$

которое можно рассматривать как кубическое уравнение относительно y с членом третьей степени, равным $4\beta y^3$, и свободным членом

$$[(\alpha-x)^2 + \beta^2]^2 - [x(1-x)]^2.$$

В критическом интервале между P_1 и P_2 эта величина становится отрицательной. Но тогда кубическое уравнение должно иметь вещественный положительный корень для всякого значения x критического интервала. Таким образом, мы получаем в верхней полуплоскости определенную выпуклую кривую, соединяющую точки P_1 и P_2 , которая вместе с отрезком вещественной оси между этими точками ограничивает область, внутри которой $A(z)$ больше единицы, а вне ее (в верхней полуплоскости) $A(z)$ меньше единицы. Выбрав наш путь интегрирования вне этой критической кривой, мы

получаем конечный путь, вдоль которого $A(z) < 1$. Тем самым мы доказали, что η_n стремится к нулю.

Единственный случай, который еще не разобран, это — равенство $\beta = 0$. Тогда z_0 лежит на оси x , и подынтегральная функция вещественна. Рассмотрим дробь

$$A(x) = \frac{x(1-x)}{(\alpha-x)^2}.$$

Ее максимум лежит в точке

$$x = \frac{\alpha}{2\alpha - 1},$$

в которой

$$A(x) = \frac{1}{4\alpha(\alpha-1)}.$$

Условие того, что эта дробь сстается меньше единицы, дает для α две границы, а именно с положительной стороны

$$\alpha > \frac{1 + \sqrt{2}}{2} = 1,20705 \quad (6-22.8)$$

и с отрицательной стороны

$$\alpha < \frac{1 - \sqrt{2}}{2} = -0,20705. \quad (6-22.9)$$

Результат нашего исследования можно резюмировать следующим образом. Удлиняем отрезок оси x симметрично по обе стороны на величину



Рис. 38.

0,207 за крайние точки 0, 1 (рис. 38). Этот отрезок окружаем оградой произвольно малой ширины. Если $f(z)$

регулярна в этой области, то сходимость квадратурной формулы обеспечена.

ЛИТЕРАТУРА К ГЛАВЕ VI

- [1] См. {11}, гл. IX.
- [1a] Никольский С. М., Квадратурные формулы, Физматгиз, М., 1958.
- [2] Lowan A. N., Davids N. and Levenson A., Table of the zeros of the Legendre Polynomials of order 1—16 and weight coefficients for Gauss' mechanical quadrature formula, Bull. Amer. Math. Soc. 48, 739 (1942).
- [3] Милн В. Э., Численное решение дифференциальных уравнений, ИЛ, М., 1956.
- [4] Сегал Б. И. и Семендяев К. А., Пятизначные математические таблицы, изд. 2, Физматгиз, 1959.

ГЛАВА VII

СТЕПЕННЫЕ РАЗЛОЖЕНИЯ

1. Историческое введение

Степени переменной x первоначально появились в чисто алгебраических задачах. В дальнейшем стало очевидным важное значение степенных разложений. Разложение, открытое Тейлором (1715) и Маклореном (1742), сделало возможным предвидение хода функции, если известно значение этой функции и всех ее производных в одной какой-нибудь точке. «Ряд Тейлора» сделался поэтому одним из краеугольных камней аналитических исследований и оказался особенно полезным для установления существования решений дифференциальных уравнений. Уже в середине XIX века стали более осторожно относиться к степенным разложениям. Простое существование разложения Тейлора еще не доказывает, что этот ряд имеет какую-нибудь внутреннюю связь с функцией, из которой он получен. С другой стороны, если ряд предназначен только для того, чтобы получить численное значение функции, то приходится исследовать характер сходимости его. Ряд Тейлора может сходиться на всей комплексной плоскости или только внутри определенного круга, но он же может всюду расходиться. Однако более свободная формулировка вопроса о сходимости значительно увеличивает пользу разложения. Можно, например, извлечь большую пользу из «полусходящихся разложений», которые фактически расходятся при возрастании числа членов до бесконечности, но сходятся в начальной стадии суммирования, позволяя, таким образом, вычислить функцию с известной ограниченной точностью; однако эта точность не может быть увеличена, так как погрешность усеченного ряда убывает до определенного минимума, а затем снова возрастает. Много внимания также уделялось нахождению методов «обобщенного суммирования» ряда: после применения этого метода ряд становится сходящимся, хотя частная сумма первоначального ряда при увеличении числа членов возрастает до бесконечности.

С развитием теории ортогональных разложений выяснилось, что иногда степенные разложения, коэффициенты которых определяются не по схеме Тейлора, могут быть гораздо более эффективными, чем

ряд Тейлора. Такие разложения основаны не на процессе дифференцирования, а на процессе интегрирования. Обширный класс функций недостаточно аналитических, чтобы их можно было разложить в ряд Тейлора, может быть представлен с помощью таких ортогональных разложений. Таким образом, область степенных разложений расширилась за пределы семейства аналитических функций. Но даже для аналитических функций мы можем выиграть в сходимости, если применить не степени переменной непосредственно, а полиномы, являющиеся членами ортогональной системы функций. Эти разложения определены в заданной конечной вещественной области переменной x , и нашей целью является аппроксимирование функции на всем интервале, а не в отдельных точках. Выигрыш по сравнению с рядом Тейлора состоит в том, что мы жертвуем точностью в той точке, где ряд Тейлора давал весьма точные результаты, но уменьшаем погрешность в периферийных областях, где ошибка разложения в ряд Тейлора делалась недопустимо большой.

Теория ортогональных разложений основана на методе наименьших квадратов, развитом Гауссом и Лежандром. Многие системы ортогональных функций стали известны в связи с изучением оператора Лапласа (уравнения потенциала), играющего фундаментальную роль не только в астрономии, но и в теории электричества и магнетизма. Ряд Фурье был первым классическим примером ортогонального разложения. Однако природа ортогональных разложений была понята лишь после замечательных исследований по акустике Релея (Rayleigh) (Теория звука, 1894); его и следует считать истинным основоположником теории ортогональных разложений. Теория интегральных уравнений Фредгольма (1900) и дальнейшие исследования Гильберта (1910) дали мощный стимул к более широкому пониманию задач ортогональности и их отношения к метрической геометрии.

В настоящей главе нас будет особенно интересовать вопрос о *быстро сходящихся степенных разложениях*. Одна лишь сходимость разложения, как бы ценна она ни была с чисто теоретической точки зрения, практически приносит мало пользы, если число членов, необходимых для достижения разумной точности, очень велико. Можно, однако, построить методы, с помощью которых длинное степенное разложение может быть «телескопически сдвинуто»*) в гораздо более короткий ряд, состоящий из нескольких членов. Конечный степенной ряд не может быть заменен другим рядом, если требуется *абсолютная* точность. Однако если нас удовлетворяет заданная ограниченная точность, скажем в 5% , то может случиться, что длинное степенное разложение, содержащее, скажем, 50 членов, может быть заменено другим разложением только из пяти членов.

*) Как трубка старинного ручного телескопа (зрительной трубы), составленная из многихдвигающихся друг в друга звеньев.

Этот «телескопический сдвиг» степенного ряда осуществляется не *опусканием* членов, начиная с некоторого, а *пересоставлением* ряда, согласно определенной схеме. Другая возможность состоит в использовании дифференциального уравнения, которое определяет некоторую функцию с тем, чтобы при решении его с помощью степенного ряда погрешности были равномерно распределены по всему интервалу. В этих исследованиях фундаментальную роль играет некоторый важный класс полиномов, введенных русским математиком Чебышевым (1864) и названных «полиномами Чебышева».

2. Аналитическое разложение с помощью обратных радиусов

Разложение в ряд Тейлора аналитической функции $w = f(z)$ имеет некоторый радиус сходимости, который определяется аналитической природой функции $f(z)$. Если $f(z)$ становится бесконечной в некоторой точке $z = z_0$ или имеет другую какую-нибудь «особенность» в этой точке, которая не согласуется с общими условиями аналитического поведения, то бесконечный ряд

$$w = a_0 + a_1 z + a_2 z^2 + \dots \quad (7-2.1)$$

не может сходиться вне круга $|z| = |z_0|$. Но ряд (1) будет сходиться *внутри* круга, если известно, что $z = z_0$ есть *ближайшая* к началу $z = 0$ точка, в которой нарушается аналитическое поведение функции $f(z)$. На самой границе круга сходимости ряд может сходиться или расходиться, и выяснение характера сходимости требует особого исследования.

Для примера рассмотрим две функции:

$$w = e^{z^2}, \quad (7-2.2a)$$

$$w = \frac{1}{1 + e^z}. \quad (7-2.2b)$$

Первая функция остается конечной и аналитической во всех точках комплексной плоскости $z = x + iy$. Следовательно, ряд Тейлора в окрестности начала должен сходиться при любом конечном значении z . Такого рода функция называется «целой функцией».

Вторая функция принимает бесконечные значения в точках

$$z = \pm \pi i, \pm 3\pi i, \pm 5\pi i, \dots$$

Ближайшие к началу особые точки суть $z = \pm \pi i$. Поэтому ряд Тейлора (2.1) для этой функции сходится лишь внутри круга $|x + iy| = \pi$.

Часто можно получить коэффициенты ряда Тейлора, исходя из определяющего уравнения для функции, не прибегая к процессу последовательного дифференцирования. Например, формула (2b) может

быть записана в виде

$$(1 + e^z) w = 1.$$

Разложив оба множителя в ряд Тейлора, получим

$$\left(2 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \dots\right)(a_0 + a_1 z + a_2 z^2 + \dots) = 1.$$

Если мы выполним умножения почленно, то получим формулы для последовательного определения коэффициентов a_i :

$$\begin{aligned} 2a_0 &= 1 & \left(a_0 = \frac{1}{2}\right), \\ 2a_1 + a_0 &= 0 & \left(a_1 = -\frac{1}{4}\right), \\ 2a_2 + a_1 + a_0 &= 0 & (a_2 = 0), \\ 2a_3 + a_2 + \frac{a_1}{2} + \frac{a_0}{6} &= 0 & \left(a_3 = \frac{1}{48}\right), \\ &\dots & \dots \end{aligned}$$

Учитывая, что наш ряд расходится вне определенного круга, можно поставить вопрос, нельзя ли каким-нибудь способом выйти за пределы круга сходимости. Большое теоретическое, но небольшое практическое значение имеет метод «аналитического продолжения». При этом методе мы выбираем точку, лежащую еще внутри круга сходимости, например для функции (2б) точку $z = \frac{2}{3}\pi$. При помощи ряда Тейлора получаем значения функции и всех ее производных в этой точке и принимаем ее за центр нового разложения Тейлора. Радиус нового круга сходимости опять определяется ближайшими особыми точками, т. е. точками $z = \pm i\pi$. При помощи нового разложения мы теперь покрываем серпообразную область, не входившую в прежний круг сходимости. Таким образом, для функции, первоначально заданной бесконечным рядом (1), мы можем распространить область определения, ограниченную сначала определенным кругом, на все большие и большие части комплексной плоскости.

Численно гораздо более удобный метод можно получить, используя «преобразование обратными радиусами». Многие важные трансцендентные функции прикладного анализа обладают тем свойством, что они являются аналитическими во всей правой комплексной полуплоскости, хотя и имеют особенность в точке $z = 0$. Допустим, что у нас имеется для такой функции разложение Тейлора в круге с центром в некоторой точке $z = p$, где p — вещественное положительное число:

$$w = f(z) = a_0 + a_1(z - p) + a_2(z - p)^2 + \dots \quad (7-2.3)$$

Это разложение сходится для всех z , расположенных внутри круга $|z| = p$.

Применим теперь преобразование

$$z = p \frac{1 + \xi}{1 - \xi}. \quad (7-2.4)$$

Оно обладает тем свойством, что вся правая полуплоскость переменной z отображается в единичный круг $|\xi| = 1$ новой переменной ξ . Следовательно, функция $w = f(z)$, рассматриваемая как функция переменной ξ , не имеет особых точек внутри единичного круга и допускает здесь разложение вида

$$w = b_0 + b_1 \xi + b_2 \xi^2 + \dots \quad (7-2.5)$$

Радиус сходимости этого ряда есть $|\xi| = 1$. Поэтому ряд (5) дает возможность вычислить $w(\xi)$ с произвольной точностью при любом значении ξ , абсолютная величина которого меньше 1, а таким значениям ξ соответствуют произвольные точки z , которые лежат в правой комплексной полуплоскости.

Но ряд (3) и ряд (5) находятся между собой в тесной связи. В самом деле,

$$z - p = 2p \frac{\xi}{1 - \xi}. \quad (7-2.6)$$

Заменим это преобразование двумя последовательными преобразованиями

$$z - p = 2pt \quad \text{и} \quad t = \frac{\xi}{1 - \xi}. \quad (7-2.7)$$

Первое преобразование дает

$$w = a_0 + 2pa_1 t + (2p)^2 a_2 t^2 + \dots = a'_0 + a'_1 t + a'_2 t^2 + \dots, \quad (7-2.8)$$

где

$$a'_k = (2p)^k a_k. \quad (7-2.9)$$

Второе преобразование означает, что член $a'_k t^k$ надо заменить бесконечным рядом

$$a'_k t^k = a'_k \frac{\xi^k}{(1 - \xi)^k} = a'_k \xi^k \sum_{\alpha=0}^{\infty} \binom{k + \alpha - 1}{\alpha} \xi^\alpha = a'_k \sum_{m=k}^{\infty} \binom{m-1}{k-1} \xi^m. \quad (7-2.10)$$

Поэтому переход от ряда (3) к ряду (5) приводит к следующим значениям для коэффициентов b_k :

$$b_{k+1} = \sum_{\alpha=0}^k \binom{k}{\alpha} a'_{k+1}, \quad (7-2.11)$$

или более подробно:

$$\left. \begin{aligned} b_0 &= a'_0, \\ b_1 &= a'_1, \\ b_2 &= a'_1 + a'_2, \\ b_3 &= a'_1 + 2a'_2 + a'_3, \\ b_4 &= a'_1 + 3a'_2 + 3a'_3 + a'_4, \\ &\dots \end{aligned} \right\} \quad (7-2.12)$$

Таким образом, коэффициенты разложения (5) можно получить путем простого суммирования коэффициентов a'_k , взятых с весами, равными соответствующим биномиальным коэффициентам.

Мы, следовательно, заменили локальное разложение (3) другим, пригодным для гораздо более обширной области. Первоначальное разложение (3) сходилось только внутри круга $|z - p| = r$ и было непригодным для вычисления $f(z)$ в какой-нибудь точке z , лежащей вне этого круга. Мы изменили коэффициенты этого разложения с помощью двух простых операций. Сначала мы умножили коэффициенты на последовательные степени числа $2r$. Затем просуммировали эти коэффициенты, приписав им веса, равные биномиальным коэффициентам. Полученные новые коэффициенты дают новый ряд, с помощью которого $f(z)$ может быть вычислена для любой точки z , лежащей в правой полуплоскости. Разложение в этом новом ряде ведется по степеням новой переменной

$$\xi = \frac{z - p}{z + p}. \quad (7-2.13)$$

Переменная ξ для всех допустимых значений z не превышает по абсолютной величине единицу.

Этот метод может быть полезным во многих задачах. Почти все основные трансцендентные функции, встречающиеся в математической физике, — такие, как функции Бесселя, функция ошибок Гаусса, гамма-функция, интегральная показательная функция — обладают тем свойством, что существенная часть функции может быть аппроксимирована простой показательной функцией, а возникающая при этом ошибка может быть скорректирована с помощью амплитудного множителя, который медленно меняется в правой комплексной полуплоскости. Этот амплитудный множитель также регулярен на всей комплексной полуплоскости всюду, за исключением особых точек $z = 0$ и $z = \infty$. Кроме того, определяющее дифференциальное или функциональное уравнение позволяет получить последовательные производные этой амплитудной функции в надлежаще выбранной точке $z = p$. Но мы имеем теперь метод, с помощью которого такая функция может быть аналитически представлена и численно определена во всех интересующих нас точках на основе этих производных.

3. Численный пример

Мы проиллюстрируем действие этого метода на простом, но типичном примере. Рассмотрим «интегральную показательную функцию»

$$w(z) = \int_z^{\infty} \frac{e^{-t}}{t} dt. \quad (7-3.1)$$

Определяющим дифференциальным уравнением здесь будет

$$w'(z) = -\frac{e^{-z}}{z}. \quad (7-3.2)$$

Интегрирование по частям показывает, что для больших значений z имеет место приближенное соотношение

$$w(z) = \frac{e^{-z}}{z}.$$

Поэтому мы положим

$$w(z) = \frac{e^{-z}}{z} u(z) \quad (7-3.3)$$

и рассмотрим неизвестную функцию $u(z)$:

$$u(z) = ze^z w(z). \quad (7-3.4)$$

Подстановка в (2) дает определяющее уравнение для $u(z)$:

$$u' - \frac{u}{z} - u = -1, \quad \text{или} \quad -zu' + (1+z)u = z.$$

Нам нужно получить разложение u в окрестности точки $p=1$. Поэтому полагаем

$$z = 1 + t, \quad -(1+t)u' + (2+t)u = 1+t. \quad (7-3.5)$$

Если подставить в это уравнение вместо u ряд

$$u = a_0 + a_1 t + a_2 t^2 + \dots \quad (7-3.6)$$

и составить таблицу для вычисления коэффициентов, то получим последовательно

$$\left. \begin{aligned} a_1 &= -1 + 2a_0, & a_5 &= \frac{1}{5}(a_3 - 2a_4), \\ a_2 &= \frac{1}{2}(-1 + a_0 + a_1), & a_6 &= \frac{1}{6}(a_4 - 3a_5), \\ a_3 &= \frac{1}{3}a_1, & a_7 &= \frac{1}{7}(a_5 - 4a_6), \\ a_4 &= \frac{1}{4}(a_2 - a_3), & & \dots \end{aligned} \right\} \quad (7-3.7)$$

Таким образом, все коэффициенты однозначно определяются, если задать a_0 . Значение коэффициента a_0 берем из таблицы¹⁾

$$a_0 = e \int_1^{\infty} \frac{e^{-t}}{t} dt = 0,596347361.$$

Последовательные подстановки дают

$$\left. \begin{array}{ll} a_1 = 0,192694722, & a_5 = 0,029817368, \\ a_2 = 0,105478958, & a_6 = -0,021979956, \\ a_3 = 0,064231574, & a_7 = 0,016819599, \\ a_4 = -0,042427633, & \dots \end{array} \right\} \quad (7-3.8)$$

Следующим шагом является умножение на последовательные степени $2p$, т. е., в нашем случае, на степени 2. Полученные коэффициенты a_i затем умножаются на веса, равные биномиальным коэффициентам:

$$\begin{array}{ll} a'_1 = 0,385389444, & 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \\ a'_2 = -0,421915834, & 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \\ a'_3 = 0,513852592, & 1 \quad 3 \quad 6 \quad 10 \quad 15 \\ a'_4 = -0,678842130, & 1 \quad 4 \quad 10 \quad 20 \\ a'_5 = 0,954155779, & 1 \quad 5 \quad 15 \\ a'_6 = -1,406717197, & 1 \quad 6 \\ a'_7 = 2,152908672, & 1 \end{array} \quad (7-3.9)$$

Если ограничиться восемью членами, то в результате получится ряд

$$u(\xi) = 0,5963 + 0,3854\xi - 0,0365\xi^2 + 0,0554\xi^3 - 0,0176\xi^4 + 0,0196\xi^5 - 0,0100\xi^6 + 0,00978\xi^7. \quad (7-3.10)$$

Этот ряд медленно сходится при приближении к единичному кругу $|\xi| = 1$. Вообще, аппроксимация этого рода не обладает особенно большой точностью, и мы располагаем другими методами для приближения интеграла показательной функции, которые дают гораздо бóльшую точность. Но эти другие методы используют дифференциальное уравнение, определяющее функцию $u(z)$, тогда как разложение обратными радиусами может быть получено при гораздо более общих условиях. Ценность этого метода — в том, что путем простого взвешенного суммирования мы получаем из локального разложения Тейлора новое разложение, которое представляет искомую функцию $u(z)$ в обширной области переменной z в виде простого аналитического выражения (умеренной точности).

¹⁾ См. ссылку в IV, 21

Максимальную ошибку разложения (10) можно оценить следующим образом. Особая точка функции $u(z)$ есть $z=0$, т. е. $\xi=-1$. Здесь $u(z)$ стремится не к бесконечности, а к нулю, как $z \log z$. Мы можем поэтому ожидать, что ряд пригоден даже на предельной окружности $|\xi|=1$, которая соответствует мнимой оси переменной z . Сходимость будет самой слабой в особой точке $z=-1$. Здесь ряд (10) дает

$$u(-1) = 0,0619.$$

Следовательно, есть основания считать, что абсолютная величина погрешности нигде не превысит 0,062.

В качестве численного опыта найдем из нашего приближения значение функции $w(z)$ в точке $z=i$. Соответствующее значение ξ есть

$$\xi = \frac{i-1}{i+1}.$$

Мы находимся, таким образом, на предельной окружности. Подставив в (10), получаем

$$u(i) = 0,6253 + 0,3398i.$$

При обратном переходе к первоначальной функции $w(z)$ по формуле (3) мы разложим $w(z)$ на интегральный косинус $ci(z)$ и интегральный синус $si(z)$. Тогда получим

$$\begin{aligned} w(i) &= -ci(1) - i\left(\frac{\pi}{2} - si(1)\right) = \frac{1}{i}(\cos 1 - i \sin 1)u(z) = \\ &= -(0,6253 + 0,3398i)(0,8415 + 0,5403i) = -0,3425 - 0,6238i, \end{aligned}$$

откуда

$$ci(1) = 0,3425, \quad si(1) = 0,9470,$$

тогда как на самом деле эти значения суть

$$ci(1) = 0,3374, \quad si(1) = 0,9461.$$

Мы, таким образом, получили точность в 1% .

4. Сходимость ряда Тейлора

Коэффициенты ряда Тейлора получаются путем последовательных дифференцирований в определенной точке $z=z_0$ (центре разложения). Так как производные дифференцируемой функции являются все пределами разностных коэффициентов, получающимися при стремлении h к нулю, то, очевидно, коэффициенты ряда Тейлора определяются однозначно, если функция известна в *произвольно малой* окрестности точки $z=z_0$. Мы приходим, таким образом, к замечательному результату, что ход аналитической функции может быть *предсказан*, если она известна только в малом. Ряд Тейлора является, следовательно, экстраполирующим, а не интерполирующим рядом. Так как интерпо-

ляция всегда надежнее, чем экстраполяция, следует ожидать, что ряд, основанный на интерполяции, дал бы благодаря более сильной сходимости гораздо лучшие результаты, чем ряд Тейлора. Хотя, как мы убедимся дальше, эта точка зрения правильна, она верна лишь тогда, когда мы интересуемся *одномерным* интервалом переменной z . Положение, однако, оказывается иным, если мы исследуем всю *двумерную* область, заключенную в круге сходимости. В этой области пригодность ряда Тейлора вытекает из интегральной формулы Коши

$$f(z) = \frac{1}{2\pi i} \oint \frac{f(t)}{t - z} dt, \quad (7-4.1)$$

где контурный интеграл берется по граничным значениям функции $f(z)$, т. е. вдоль предельной окружности. Отсюда следует, что значения функции $f(z)$ известны *внутри* круга, если заданы ее значения *на* граничной окружности. Только благодаря аналитической природе функции $f(z)$ этот интеграл может быть заменен бесконечным сходящимся степенным рядом, коэффициенты которого получаются последовательным дифференцированием функции $f(z)$ в центре круга. Если записать $z - z_0$ в полярной форме

$$z - z_0 = re^{i\theta}, \quad (7-4.2)$$

то ряд Тейлора принимает вид

$$f(z) = \sum_{k=0}^{\infty} a_k r^k e^{ik\theta}. \quad (7-4.3)$$

Если рассматривать только значения $f(z)$ вдоль окружности определенного радиуса $r = r_0$, то $f(z)$ принимает форму ряда Фурье относительно переменной θ . Но этот ряд является уже ортогональным разложением, которое аппроксимирует функцию $f(z)$ «наилучшим» образом в смысле теории наименьших квадратов. Поэтому мы не можем ожидать улучшения сходимости ряда Тейлора, если нашей целью является получение разложения, пригодного повсюду внутри круга. При условии, что $f(z)$ должна быть аппроксимирована полиномом n -й степени, где n — заданное целое число, и что это разложение должно быть пригодным для заданной круговой области, мы не можем найти ряда, который давал бы меньшие погрешности по всей заданной области, чем усеченный ряд Тейлора.

5. Жесткие и гибкие разложения

Большое аналитическое значение ряда Тейлора и удивительные свойства ряда Фурье привели к значительной переоценке определенного рода предельного процесса. В нем мы рассматриваем бесконечный ряд следующего вида:

$$f(x) = a_1 \varphi_1(x) + a_2 \varphi_2(x) + \dots \quad (7-5.1)$$

Смысл этого «многоточия» непосредственно не уясняется, так как сложение бесконечного числа членов не может пониматься буквально. По молчаливому соглашению мы предполагаем, что ни знак «равняется», ни суммирование в правой части не следует понимать в алгебраическом смысле. Правая сторона указывает на некоторый последовательный процесс, тогда как знак равенства имеет в виду определенное соотношение между этим процессом и данной функцией $f(x)$. Процесс этот состоит в следующем. Возьмем n членов ряда для данного значения x :

$$s_n = a_1 \varphi_1(x) + a_2 \varphi_2(x) + \dots + a_n \varphi_n(x). \quad (7-5.2)$$

Затем добавим еще один член

$$s_{n+1} = s_n + a_{n+1} \varphi_{n+1}(x).$$

Далее опять добавим еще один член

$$s_{n+2} = s_{n+1} + a_{n+2} \varphi_{n+2}(x).$$

Продолжаем таким образом, постоянно добавляя еще один член. Этим определяется процесс для любого выбранного значения x .

Знак равенства между левой и правой частью выражает два независимых факта:

1. Суммы $s_n, s_{n+1}, s_{n+2}, \dots$ стремятся к определенному пределу. Это значит, что существует некоторое определенное число $l(x)$, к которому $s_n(x)$ приближается таким образом, что для достаточно большого числа n разность между $l(x)$ и $s_n(x)$ может быть сделана по абсолютной величине сколь угодно малой, хоть и не равной, вообще говоря, нулю. $l(x)$ называется «пределом сумм $s_n(x)$ при бесконечном возрастании числа n ».

2. В некотором заданном интервале переменной x функция $l(x)$ совпадает с $f(x)$.

Хотя такого рода бесконечный предельный процесс очень ценен и получил исторически господствующее значение, он тем не менее характеризуется одним серьезным недостатком, который значительно связывает нас. Переходя от n к $n+1, n+2, \dots$, мы постоянно прибавляем дополнительный член, *ничего не изменяя в сумме, которую мы уже раньше получили*. Это налагает на нас при прибавлении члена $a_k \varphi_k(x)$ тяжелую ответственность, так как мы имеем в виду использовать этот член для образования не только $s_k(x)$, когда мы впервые с ним встретились, но и для образования *всех последующих* $s_{k+\alpha}(x)$. Можно себе представить, что мы могли бы гораздо больше выиграть в эффективности приближений, если бы могли при переходе от n к $n+1$ производить две операции: добавлять еще один член и *изменять все ранее полученные коэффициенты*. В этом случае одномерная «затвердевшая» последовательность

коэффициентов

$$a_0, a_1, \dots, a_n, \dots$$

была бы заменена *двумерной* последовательностью, ибо теперь коэффициенты разложения (1) будут зависеть от n , т. е. от числа используемых членов. Таким образом, мы получаем матричную схему следующей структуры:

$$\begin{array}{ccc} a_{11} & & \\ a_{21} & a_{22} & \\ a_{31} & a_{32} & a_{33} \\ \cdot & \cdot & \cdot \end{array} \quad (7-5.3)$$

Теперь сумма $s_n(x)$ должна быть образована следующим образом:

$$s_n(x) = a_{n1}\varphi_1(x) + a_{n2}\varphi_2(x) + \dots + a_{nn}\varphi_n(x). \quad (7-5.4)$$

Как и раньше, мы переходим от n к $n+1$, $n+2$, $\dots$, и снова нашей целью является приближение $s_n(x)$ к $f(x)$ со всё возрастающей точностью. И снова мы можем сказать, что

$$f(x) = \lim_{n \rightarrow \infty} s_n(x). \quad (7-5.5)$$

Это утверждение тождественно с утверждением, составляющим содержание равенства (1), однако теперь мы уже не располагаем простой схематической записью, с помощью которой мы могли бы символически выразить предыдущие утверждения.

Использование таких «гибких» коэффициентов очень заметно увеличивает силу бесконечных разложений. Подгоняя коэффициенты к числу членов, мы можем снизить погрешность конечного ряда при том же числе членов. Кроме того, мы можем получить бесконечные ряды для функций, которые не могут быть разложены с помощью жестких коэффициентов.

Такие разложения уже применялись нами, когда мы имели дело с дифференцированием ряда Фурье (ср. IV, 6).

Мы видели, насколько полезно сгладить колебания Гиббса при помощи замены $f(x)$ функцией $\overline{f}(x)$, полученной составлением арифметической средней от функции $f(x)$ между точками $x - \frac{\pi}{n}$ и $x + \frac{\pi}{n}$.

Шаг этого процесса интегрирования изменяется с числом n . Следовательно, функция, которую мы разлагаем, постоянно изменяется с числом членов, и таким образом, коэффициенты разложения также меняются. Тем не менее при бесконечном возрастании n функция $\overline{f}_n(x)$ все более и более приближается к $f(x)$ точно так же, как и ряд Фурье, который все ближе и ближе подходит к $\overline{f}_n(x)$. Такая функция, как $y = \operatorname{tg} x$, которая не интегрируема и поэтому не может быть разложена по методу «жестких коэффициентов», тотчас же может быть разложена, если воспользоваться гибкими коэффициентами.

Однако главное значение гибких коэффициентов для парактических¹⁾ целей коренится в том обстоятельстве, что коэффициенты ортогонального разложения требуют вычисления некоторых определенных интегралов, которые часто не берутся в конечном виде. Поэтому весьма ценным обстоятельством является то, что такое разложение из n членов может быть заменено другим n -членным разложением, когда используются те же n функций, но с другими коэффициентами, которые легче вычислить, причем погрешность такого видоизмененного разложения может быть не намного большей погрешности классического разложения с коэффициентами, практически недоступными.

6. Разложение по ортогональным полиномам

В V, 20 мы рассматривали важное семейство полиномов, так называемых полиномов Якоби. Эти системы полиномов дважды бесконечного семейства отличаются друг от друга той весовой функцией, относительно которой имеет место их ортогональность. Однако при любом (положительном) весовом множителе произвольная функция, ограниченная очень немногими условиями (при пользовании методом суммирования Фейера достаточна лишь абсолютная интегрируемость; ср. IV, 2), может быть разложена по этим полиномам; таким образом, они все обладают одинаковыми аналитическими свойствами. Существует ли какое-нибудь основание, по которому мы должны бы сделать выбор в пользу одного из этих весовых множителей, раз все эти полиномы обладают одинаковой способностью представить произвольные функции?

Различия между ними обнаруживаются сами собой сразу же, если глубже вникнуть в смысл слова «представить». Легко можно ошибиться, приписав понятию «предел» некоторое магическое значение, и подумать, что выражение «пределом бесконечного ряда является $f(x)$ » означает, что надлежащая комбинация данных функций в конце концов дает $f(x)$. В действительности это «в конце концов» никогда не наступает, ибо нельзя сложить бесконечное число членов. Это выражение означает лишь то, что, добавляя все больше и больше членов, мы можем приблизиться к $f(x)$ *сколь угодно близко*. Следовательно, мы заранее соглашаемся с погрешностью и стремимся лишь к тому, чтобы эта погрешность могла быть снижена до произвольно малой величины.

Теперь, если будут заранее указаны произвольно малые пределы допустимой погрешности $\pm \epsilon$, то чистый аналитик считает достаточным показать, что сумма достаточно большого числа членов, образованных согласно определенной схеме, удовлетворит требуемому условию. Он может быть щедрым относительно числа членов, и для него

¹⁾ Значение этого термина пояснено во Введении.

безразлично, потребуется ли для этой цели 20 или 10 000 членов. Точка зрения прикладного аналитика совершенно другая. Если заданы некоторые пределы погрешности $\pm \epsilon$, то его стремлением будет сделать как можно меньшим число членов, с помощью которых результат получится с заданной точностью.

Именно в этом отношении резко отличается поведение различных классов ортогональных полиномов. Вполне может случиться, что при некотором выборе полиномов можно с двадцатью членами достичь такой точности, для которой при другого рода полиномах может потребоваться 10 000 членов. Это становится понятным, если проследить роль весового множителя $\rho(x)$. Допустим, что функция $\rho(x)$ такова, что она принимает большие значения в непосредственной окрестности нуля, но затем быстро уменьшается до почти пренебрежимо малой величины. Это значит, что мы придаем серьезное значение сходимости в непосредственной окрестности нуля за счет остальной части интервала. При этом сумма конечного числа членов такого разложения обнаружит следующее свойство. Вблизи нуля она будет давать $f(x)$ с большой точностью, но в остальной части интервала точность будет гораздо меньше. Так как, однако, предел погрешности $\pm \epsilon$ должен означать, что отклонение ряда от $f(x)$ не должно превышать $\pm \epsilon$ в *любой точке интервала* $(-1, +1)$, мы сразу видим, что неблагоприятный характер весового множителя вынудит нас набрать большое число членов, прежде чем требуемая точность ϵ будет достигнута в обоих концах ± 1 интервала.

Разложение Тейлора в окрестности точки $x=0$ фактически действует именно таким неэкономичным образом, так как всю нашу информацию относительно значений функции мы берем близ точки $x=0$. Это объясняет, почему ряд Тейлора, несмотря на его огромное аналитическое значение, часто оказывается мало полезным в практических задачах и может быть заменен гораздо более эффективными разложениями.

Остановим свое внимание на полиномах Якоби. Мы имеем дважды бесконечное семейство полиномов [ср. (5-20.5)], характеризующееся двумя параметрами γ и δ . Однако нас редко будет интересовать «кривобокое» взвешивание, при котором левая и правая части интервала $[-1, +1]$ имеют разные веса. Если мы ограничиваемся симметрическим множителем веса, то полиномы Якоби сводятся сразу к «ультрасферическим полиномам» (5-20.7), зависящим лишь от одного параметра γ . Этот параметр может все же меняться от 0 до ∞ . Случай $\gamma = \infty$ соответствует вышеупомянутому предпочтению точки $x=0$ остальным точкам и таким образом приводит к ряду Тейлора. Случай $\gamma = 1$ приводит к полиномам Лежандра, при которых осуществляется наиболее естественное взвешивание $\rho(x) = 1$. Здесь ни одной точке интервала $[-1, +1]$ не дается предпочтение. И все же это не обязательно оказывается наилучшим выбором.

Если надо исследовать погрешность усеченного ряда, мы не сильно ошибемся, оценив эту погрешность первым опущенным членом

ряда. Разложим $f(x)$ по полиномам Лежандра $P_k(x)$. Тогда погрешность $\eta_n(x)$ конечного разложения в n членов будет приближенно равна

$$\bar{\eta}_n(x) = c_n P_n(x).$$

Каков характер функции $P_n(x)$? Это — функция, которая колеблется около нуля с изменяющимися амплитудами. Амплитуды постоянно (хотя и медленно) *возрастают* при переходе через середину интервала и достигают максимума 1 в обоих конечных точках $x = \pm 1$. Следовательно, погрешность не вполне одинаково распределена по всему интервалу. Для обеспечения того, что η_n не превысит предела $\pm \varepsilon$ даже в крайних точках ± 1 , нам придется расплатиться слишком большим числом членов, ибо мы получаем *больше*, чем нам нужно внутри интервала. Не возникнет преимущества, если амплитуды колебаний погрешности будут постоянно убывать при переходе от $x = 0$ к $x = 1$. В этом случае мы получим больше, чем нужно на *краю* интервала и нам опять придется за это расплачиваться¹⁾.

Но ультрасферические полиномы $P_n^\gamma(x)$ обладают следующим свойством. Если γ имеет какое-либо значение, заключенное между ∞ и $\frac{1}{2}$, амплитуды последовательных колебаний все время *возрастают* при переходе от $x = 0$ к $x = 1$; если γ имеет значение, заключенное между $\frac{1}{2}$ и 0, они все время *убывают*. Поэтому граничное значение $\gamma = \frac{1}{2}$ особенно интересно. Здесь амплитуды остаются *постоянными* на всем интервале. Следовательно, погрешности оказываются распределенными наивыгоднейшим образом, а именно равномерно по всему интервалу. Число членов, которые при этом нужны для достижения точности $\pm \varepsilon$, не может быть уменьшено. Мы получили максимально хорошую сходимость, поскольку мы достигли наименьшего числа членов, при котором приближение функции $f(x)$ уклоняется от истинного ее значения на величину, не большую чем $\pm \varepsilon$, в любой точке заданного интервала $[-1, +1]$. Эти полиномы, называемые по имени русского математика Чебышева, являются самой элементарной (и исторически старейшей) системой полиномов ввиду их простой связи с элементарными тригонометрическими функциями $\cos n\theta$.

7. Полиномы Чебышева

Кроме наилучшей сходимости, полиномы $T_k(x)$ [ср. (5-20.12)] обладают многими другими ценными свойствами, делающими их особенно интересным классом функций. Наиболее ценное их свойство составляет тот факт, что они могут быть выражены через элементарные тригонометрические функции. Они представляют собой не что

¹⁾ Ср. [4], стр. VIII.

инное, как простые тригонометрические функции $\cos k\theta$, только выраженные через переменную

$$x = \cos \theta. \quad (7-7.1)$$

Рассмотрим произвольную интегрируемую функцию ограниченной вариации $y = f(x)$. Подставляя $\cos \theta$ вместо x , мы получаем функцию

$$y = f(\cos \theta), \quad (7-7.2)$$

которая является теперь периодической функцией переменного угла θ в интервале $[-\pi, +\pi]$. Так как $f(\cos \theta)$ есть четная функция от θ , то ее можно разложить в ряд Фурье по косинусам:

$$f(\cos \theta) = \frac{1}{2} a_0 + a_1 \cos \theta + a_2 \cos 2\theta + \dots, \quad (7-7.3)$$

в котором

$$a_k = \frac{2}{\pi} \int_0^\pi f(\cos \theta) \cos k\theta d\theta. \quad (7-7.4)$$

Этот же ряд, рассматриваемый относительно переменной x , имеет вид

$$f(x) = \frac{1}{2} a_0 + a_1 T_1(x) + a_2 T_2(x) + \dots, \quad (7-7.5)$$

где

$$a_k = \frac{2}{\pi} \int_{-1}^{+1} f(x) T_k(x) \frac{dx}{\sqrt{1-x^2}}. \quad (7-7.6)$$

Это — частный случай разложения по ультрасферическим полиномам (см. § 6) при выборе множителя $\gamma = \frac{1}{2}$, который характеризует полиномы Чебышева.

Разложение функции по полиномам Чебышева является, таким образом, просто другой интерпретацией разложения четной функции в ряд по косинусам. Это фундаментальное соотношение, которое переносит важнейшие свойства рядов Фурье в область степенных рядов, является наиболее важным свойством чебышевских полиномов.

Для численных приложений очень удобно то обстоятельство, что коэффициенты полиномов Чебышева суть *целые числа*, которые находятся в простом соотношении с биномиальными коэффициентами $\binom{n}{m}^*$. Тригонометрическая формула

$$\cos(k+1)\theta + \cos(k-1)\theta = 2 \cos \theta \cos k\theta$$

*) В русской литературе обозначают биномиальный коэффициент $\binom{n}{m}$ символом C_m^n .

дает при переходе к переменной x рекуррентную формулу для полиномов Чебышева ¹⁾

$$T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x) \quad (7-7.7)$$

с начальными условиями

$$T_0(x) = 1, \quad T_1(x) = x. \quad (7-7.8)$$

Это приводит все время к целым коэффициентам. В действительности, формула (7) показывает, что коэффициент при x^m должен и при делении на 2^{m-1} оставаться целым. Если мы запишем $T_k(x)$ в виде

$$T_k(x) = c_k^0 + c_k^1 x + c_k^2 x^2 + \dots + c_k^k x^k, \quad (7-7.9)$$

то для коэффициентов c_k^m получим выражение

$$c_k^m = 2^{m-1} \left[2 \binom{\frac{1}{2}(k+m)}{\frac{1}{2}(k-m)} - \binom{\frac{1}{2}(k+m)-1}{\frac{1}{2}(k-m)} \right] (-1)^{\frac{k-m}{2}}, \quad (7-7.10)$$

помня при этом, что коэффициент, для которого $k+m$ есть нечетное число, считается равным нулю. Первые десять полиномов $T_k(x)$ даны в приложении (табл. V).

8. Смещенные полиномы Чебышева

Во многих задачах прикладного анализа нормирование середины интервала к точке $x=0$ оказывается неудобным. Часто возникает необходимость разложения функции в окрестности точки $x=0$, но при этом нас интересуют лишь положительные значения x . Иногда нам может понадобиться разложение в окрестности точки $x=\infty$; используя преобразование обратными радиусами и вводя новую переменную $\xi = \frac{1}{x}$, мы сводим эту задачу к предыдущей. Здесь снова аналитические свойства $f(\xi)$ слева и справа от $\xi=0$ обыкновенно оказываются совершенно несходными и не могут рассматриваться вместе. Поэтому нормирование интервала к $[0, 1]$ часто бывает гораздо более удобным, чем прежнее нормирование к интервалу $[-1, 1]$. Это, в частности, относится к случаю, когда мы хотим воспользоваться «параметрическим методом», с которым мы встретились в гл. VI, 21 [см. (6-21.12)]. Положим

$$x = \alpha x_1 \quad (7-8.1)$$

¹⁾ По поводу дальнейших алгебраических и функциональных свойств многочленов $T_k(x)$ см. [4], Введение.

и будем считать точку $x = \alpha$ соответствующей точке $x_1 = 1$ новой переменной. Хотя при решении задачи α рассматривается как постоянная, она может быть отождествлена с любым комплексным значением z . Поэтому, взяв значение функции в конечной точке $x_1 = 1$ интервала, мы фактически получаем $f(z)$ относительно первоначальной переменной. Здесь снова требуется, чтобы x_1 был нормирован к интервалу $[0, 1]$.

Преобразование полинома $T_k(x)$ к новому интервалу оказывается весьма простым. Мы теперь полагаем

$$\cos \theta = 2x - 1, \quad (7-8.2)$$

или

$$x = \frac{1 + \cos \theta}{2} = \cos^2 \frac{\theta}{2}. \quad (7-8.3)$$

Когда θ изменяется от 0 до π , x изменяется от 1 до 0. Полиномы Чебышева снова определяются равенством

$$T_k^*(x) = \cos k\theta, \quad (7-8.4)$$

но теперь уже относительно новой переменной (3). Новые полиномы имеют совершенно другие коэффициенты; они могли бы быть получены из коэффициентов прежних полиномов $T_k(x)$ путем простой подстановки

$$T_k^*(x) = T_k(2x - 1). \quad (7-8.5)$$

Этот полином $T_k^*(x)$ теперь просто равен с точностью до знака гипергеометрической функции [ср. (5-20.12)]

$$T_k^*(x) = (-1)^k F\left(-k, k, \frac{1}{2}; x\right). \quad (7-8.6)$$

Если опять записать $T_k^*(x)$ в форме (7.9), то коэффициенты c_k^m теперь будут

$$c_k^m = 2^{2m-1} \left[2 \binom{k+m}{k-m} - \binom{k+m-1}{k-m} \right] (-1)^{m+k}. \quad (7-8.7)$$

Эти полиномы уже не являются попеременно четными и нечетными функциями. Они содержат коэффициенты *всех* степеней между 0 и k с чередующимися знаками $\pm$. Первые 10 полиномов T_k^* приведены в приложении (табл. VII).

Полиномы $T_n^*(x)$ определены формулами (4), (3). Согласно этому общему определению мы должны положить $T_0^*(x) = 1$. Однако оказывается, что во многих формулах суммирования, в которые входят полиномы Чебышева, полином $T_0^*(x)$ входит только с половинным

весом и должен учитываться отдельно. Чтобы избежать этого неудобства, мы установим следующее соглашение. Хотя, согласно (4), (3), $T_0^*(x) = 1$, мы полагаем

$$T_0^* = \frac{1}{2}. \quad (7-8.8)$$

9. Телескопический сдвиг степенного ряда путем последовательного сокращения

Из рассуждений § 6 видно, что одна и та же функция может быть разложена в целый спектр различных степенных рядов. Все они представляют одну и ту же функцию, но обладают весьма различной скоростью сходимости. Если нашей целью является абсолютная точность, то все эти разложения по существу равнозначны. Нам нужно вычислить все большее и большее число членов, чтобы сделать погрешность меньшей, чем произвольно заданная малая величина ϵ . Если, однако, нашей целью является заданная ограниченная точность, эти разложения будут совершенно различны. Мы можем произвести разложение непосредственно по степеням x , как это мы делаем в случае ряда Тейлора. В этом случае сходимость оказывается самой слабой, т. е. погрешность, если оборвать ряд после n -го члена, оказывается наибольшей. На другом конце спектра находятся полиномы Чебышева. Если разлагать по этим полиномам, то сходимость — самая сильная, т. е. погрешность, если оборвать ряд после n -го члена, будет самой малой. Действительно, мы получили численное сравнение обоих разложений и нашли, что во втором из них порядок величины погрешности уменьшается в 2^{n-1} раз (в случае интервала $[0,1]$ в 2^{2n-1} раз). Это дает возможность получить большую точность даже с малым числом членов.

Лишь очень редко удастся получить разложение по полиномам Чебышева путем вычисления коэффициентов a_k по формуле (7.6). Вычисление определенных интегралов (7.6), как правило, окажется чересчур громоздким. Но существуют другие приемы, которые зачастую требуют гораздо меньше труда. Мы можем, например, начать с ряда Тейлора, часто имеющегося в нашем распоряжении, и «телескопически сдвинуть» его в гораздо более короткий ряд, не теряя существенно в точности. Чтобы наглядно показать действие этого метода, допустим, что задан некоторый конечный полином n -й степени. Мы ставим вопрос: нельзя ли заменить этот полином другим, меньшей степени, без значительной потери точности?

Предположим, что задана следующая функция в интервале $[0, 1]$:

$$y = 1 - x + x^2 - x^3 + x^4 - x^5 + x^6. \quad (7-9.1)$$

График этой функции вполне гладкий (рис. 39). Он соответствует полиному шестой степени. Последним членом никак нельзя пренебречь,

не совершив большой ошибки. Поэтому для представления функции y нужны, казалось бы, все степени, от первой до шестой, и все же это не так. Надлежащим приемом мы можем упростить наше длинное разложение.

Выпишем определяющее равенство для $T_6^*(x)$ в следующем виде:

$$x^6 = \frac{6144x^5 - 6912x^4 + 3584x^3 - 840x^2 + 72x - 1}{2048} + \frac{T_6^*(x)}{2048}. \quad (7-9.2)$$

Это равенство обнаруживает замечательный факт. Хотя член x^6 алгебраически и не зависит от низших степеней и, конечно, не может быть выражен линейной комбинацией низших степеней, однако эта независимость, если рассматривать интервал $[0, 1]$, удивительно слаба. В этом интервале степень x^6 почти равна определенной линейной комбинации низших степеней. Если мы положим

$$x^6 = \frac{6144x^5 - 6912x^4 + 3584x^3 - 840x^2 + 72x - 1}{2048}, \quad (7-9.3)$$

т. е.

$$x^6 = 3x^5 - 3,375x^4 + 1,75x^3 - 0,410156x^2 + 0,035156x - 0,000488, \quad (7-9.4)$$

то совершим ошибку, которая нигде не превышает $\frac{1}{2048} = 0,000488$, в силу того обстоятельства, что $T_6^*(x)$ на всем интервале колеблется

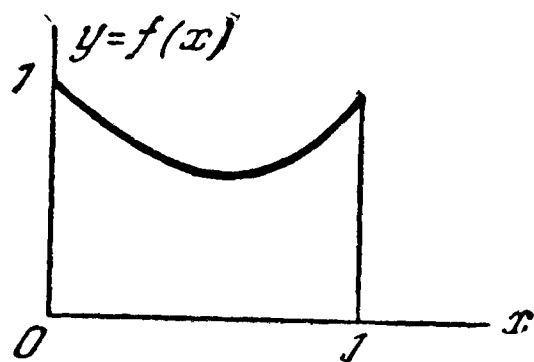


Рис. 39.

между значениями ± 1 . Кроме того, эта точность не может быть превзойдена, ибо коэффициент 2048 в $T_6^*(x)$ есть наибольший возможный коэффициент какого-либо полинома, который в данном интервале колеблется между ± 1 .

Как видим, если мы не пренебрежем членом x^6 , а заменим его линейной комбинацией низших степеней согласно равенству (4), то совершим ошибку, которая очень мала и вполне может оказаться ниже требуемой точности. Подставив выражение (4) в заданную функцию $y = f(x)$, получаем приближение

$$\bar{y} = 0,999512 - 0,964844x + 0,589844x^2 + 0,75x^3 - 2,375x^4 + 2x^5 \quad (\pm 0,0005). \quad (7-9.5)$$

Новая функция y есть полином лишь *пятой* степени.

Мы можем, очевидно, продолжить этот процесс. Табл. VII полиномов Чебышева $T_k^*(x)$ показывает, что x^5 может быть следующим

образом выражена через низшие степени:

$$\begin{aligned} x^5 &= \frac{1280x^4 - 1120x^3 + 400x^2 - 50x + 1}{512} \left(\pm \frac{1}{512} \right) = \\ &= 2,5x^4 - 2,1875x^3 + 0,78125x^2 - 0,09765x + 0,001953 \\ &\quad (\pm 0,00195). \end{aligned} \quad (7-9.6)$$

Подставив это выражение в полученное раньше выражение $\bar{y}$, мы получаем новое приближение, в котором используются только *четыре* степени:

$$\begin{aligned} \bar{y} &= 1,003418 - 1,160156x + 2,152344x^2 - 3,625x^3 + 2,625x^4 \\ &\quad (\pm 0,0044). \end{aligned} \quad (7-9.7)$$

Ошибка несколько возросла, но она может все же удовлетворять требованиям точности. Поэтому мы можем снова попытаться счастья. Согласно табл. VII x^4 может быть выражен через низшие степени с точностью, которая несколько меньше 0,01:

$$\begin{aligned} x^4 &= \frac{256x^3 - 160x^2 + 32x - 1}{128} \left(\pm \frac{1}{128} \right) = \\ &= 2x^3 - 1,25x^2 + 0,25x - 0,007812 (\pm 0,00781). \end{aligned} \quad (7-9.8)$$

Подставляя снова в (9.7), мы теперь получаем приближение

$$\bar{y} = 0,982910 - 0,503906x - 1,128906x^2 + 1,625x^3 (\pm 0,0249). \quad (7-9.9)$$

Теперь ошибка гораздо больше, чем раньше, но все же достаточно мала для большинства практических целей.

Если сделать еще один шаг дальше, можно подставить

$$\begin{aligned} x^3 &= \frac{48x^2 - 18x + 1}{32} \left(\pm \frac{1}{32} \right) = \\ &= 1,5x^2 - 0,5625x + 0,03125 (\pm 0,03125) \end{aligned} \quad (7-9.10)$$

и получить

$$\bar{y} = 1,0337 - 1,4180x + 1,3086x^2 (\pm 0,0757). \quad (7-9.11)$$

Первоначальный длинный полином шестой степени сведен к новому полиному лишь второй степени при точности, которая может оказаться достаточной во многих случаях.

10. Телескопический сдвиг степенного ряда путем пересоставления

Если мы проанализируем подробно те действия, которые мы произвели в предыдущем параграфе, то придем к новой формулировке нашего процесса. Предположим, что мы ничего не опускаем, но записываем первый шаг процесса сокращения в виде

$$y = y_5 + 0,000488T_6^*.$$

Следующий шаг тогда может быть записан так:

$$y_5 = y_4 + 0,003906T_5^*.$$

Дальнейший шаг дает

$$y_4 = y_3 + 0,020508T_4^*$$

и последний шаг —

$$y_3 = y_2 + 0,050781T_3^*.$$

Объединение всех этих шагов дает

$$y = y_2 + 0,050781T_3^* + 0,020508T_4^* + 0,003906T_5^* + 0,000488T_6^*.$$

Предположим, что мы продолжили этот процесс сокращения до самого конца, придя в конце концов к постоянной. Мы получили бы [при $T_0^* = \frac{1}{2}$, ср. (8.8)]

$$y = 1,630859T_0^* - 0,054687T_1^* + 0,163574T_2^* + 0,050781T_3^* + \\ + 0,020508T_4^* + 0,003906T_5^* + 0,000488T_6^*.$$

То, что мы здесь получили, уже не является обыкновенным степенным рядом, а *разложением по полиномам Чебышева*. Конечно, перед нами та же самая функция, из которой мы исходили. Лишь форма, в которой она представлена, другая. Однако это изменение формы исключительно благоприятно для парактических целей.

Вместо того чтобы выполнять шаг за шагом, мы можем действовать более систематически следующим образом. Точно так же как $T_n^*(x)$ выражаются в виде линейных комбинаций степеней x , мы можем обернуть процесс и выразить степени переменной x в виде линейных комбинаций полиномов $T_n^*(x)$. Это легко выполнить, если вспомнить первоначальное определение полиномов T_n^* :

$$T_n^*(x) = \cos n\theta, \quad x = \cos^2 \frac{\theta}{2}. \quad (7-10.1)$$

Отсюда

$$x^n = \cos^{2n} \frac{\theta}{2} = \left(\frac{e^{i\frac{\theta}{2}} + e^{-i\frac{\theta}{2}}}{2} \right)^{2n} = \frac{2}{4^n} \left[\cos n\theta + \binom{2n}{1} \cos (n-1)\theta + \right. \\ \left. + \binom{2n}{2} \cos (n-2)\theta + \dots + \binom{2n}{n} \frac{1}{2} \right] = \\ = \frac{2}{4^n} \left[T_n^*(x) + \binom{2n}{1} T_{n-1}^*(x) + \right. \\ \left. + \binom{2n}{2} T_{n-2}^*(x) + \dots + \binom{2n}{n} T_0^* \right]. \quad (7-10.2)$$

Отсюда видно, что степени x^n , если не считать постоянного множителя, оказываются очень простой биномиально взвешенной суммой полиномов $T_k^*(x)$.

Мы можем, следовательно, присоединить к таблице полиномов Чебышева $T_k^*(x)$ другую таблицу, решающую обратную задачу; она выражает степени x через полиномы $T_i^*(x)$ (ср. табл. VIII приложения), например,

$$1 = 2T_0^*,$$
$$x = \frac{1}{2}(2T_0^* + T_1^*),$$
$$x^2 = \frac{1}{8}(6T_0^* + 4T_1^* + T_2^*),$$
$$\dots\dots\dots$$

Этот процесс обращения численно не слишком сложен, в особенности если умножить ряд (2) на надлежащий множитель, который устраняет все знаменатели. Остальное представляет собой только умножение на коэффициенты и сложение, что может быть выполнено на арифмометре путем процесса накопления. Например, для решения задачи § 9 мы умножим заданные коэффициенты, в порядке возрастающих степеней, на множители

$$\frac{4096, -1024, 256, -64, 16, -4, 1}{2048}.$$

Это дает, если опустить знаменатель,

$$4096, -1024, 256, -64, 16, -4, 1.$$

Теперь эта строка умножается на последовательные столбцы следующей матрицы, которая берется непосредственно из табл. VIII ¹⁾:

1						
2	1					
6	4	1				
20	15	6	1			
70	56	28	8	1		
252	210	120	45	10	1	
924	792	495	220	66	12	1

Это дает

$$2048y = 3340T_0^* - 112T_1^* + 335T_2^* + 104T_3^* + 42T_4^* + 8T_5^* + T_6^*.$$

¹⁾ Недостающие элементы матрицы предполагаются равными нулю.

Отсюда, разделив на 2048, получаем

$$y \Rightarrow 1,630859T_0^* - 0,054687T_1^* + 0,163574T_2^* + 0,050781T_3^* + \\ + 0,020508T_4^* + 0,003906T_5^* + 0,000488T_6^*.$$

Этой перестановкой членов первоначального степенного ряда мы добиваемся усиления сходимости ряда. Коэффициентами первоначального ряда были

$$1, -1, 1, -1, 1, -1, 1.$$

Коэффициенты нового ряда имеют значения

$$1,6308; -0,0547; 0,1636; 0,0508; 0,0205; 0,0039; 0,0005.$$

Если не опускать членов, новый ряд представляет собой просто видоизмененную форму первоначального ряда и никакого преимущества не дает. В действительности, однако, гораздо более быстрая сходимость нового ряда дает нам возможность *опустить некоторые члены*. Погрешность, вызванная опущенными членами, дается просто их коэффициентами. Мы получаем верхнюю границу ошибки в любой точке интервала, складывая абсолютные значения коэффициентов всех опущенных членов. В нашем примере все последние коэффициенты случайно оказались положительными числами. Это, однако, совершенно несущественно, так как ошибка, допущенная при опускании какого-нибудь члена, всегда имеет меняющийся знак, и поэтому для реалистической оценки максимальной погрешности обязательно следует брать *сумму абсолютных значений всех опущенных коэффициентов*.

В нашей задаче, например, мы получаем следующие последовательные оценки погрешности:

$$0,0005; 0,0044; 0,0249; 0,0757.$$

В зависимости от желаемой точности мы сразу можем сказать, сколько членов можно опустить, совершив ошибку, не превышающую допустимый предел.

Следует ли предпочесть процесс последовательного сокращения, изложенный в прошлом параграфе, процессу систематической перестановки настоящего параграфа — это зависит от численного характера задачи. Оба процесса ведут к одному и тому же результату, а именно к *телескопическому сдвигу степенного разложения*, создающему максимально улучшенную сходимость.

11. Степенные разложения вне интервала сходимости Тейлора

Методы двух предыдущих параграфов применимы лишь тогда, когда мы исходим из некоторого уже заданного степенного разложения. Мы нашли прием, с помощью которого этот ряд может быть заменен более коротким рядом, если можно допустить некоторую

погрешность. Мы могли бы, следовательно, начать, скажем, с заданного разложения Тейлора и привести его к более экономному разложению, содержащему меньше членов. Но эта процедура может оказаться весьма громоздкой при приближении точки интервала к границе круга сходимости ряда Тейлора, так как тогда нам надо было бы брать с самого начала очень большое число членов, и численные выкладки могли бы стать неодолимо трудными.

Еще хуже положение, когда нужно получить разложение в степенной ряд в интервале, выходящем за пределы сходимости ряда Тейлора. Например, функция

$$y = \frac{1}{1+x} \quad (7-11.1)$$

разлагается в сходящийся ряд Тейлора только вплоть до $x=1$, тогда как нам может понадобиться аппроксимировать эту функцию в интервале $[0, 4]$. Здесь процесс перегруппировки теряет свое значение, так как нет никакой гарантии, что, оперируя с расходящимся рядом, мы придем к сходящемуся результату.

Необходимо поэтому найти методы, которые давали бы практически хорошо сходящиеся разложения, даже если мы не можем опереться на какой-нибудь исходный ряд, который может быть перегруппирован к более эффективному виду. Существование таких рядов гарантируется теорией ортогональных разложений, в которой доказывается, что любая квадратично интегрируемая функция ограниченной вариации может быть разложена по функциям полной ортогональной системы, таких, как полиномы Лежандра или Чебышева. Следовательно, в возможности разложить заданную функцию $y=f(x)$ вне интервала сходимости ряда Тейлора не может быть сомнений; трудность состоит лишь в том, что мы не располагаем коэффициентами (7.6) такого разложения, ибо не знаем, как практически вычислить определенные интегралы, которыми определяются эти коэффициенты. Нужно найти другие удобные методы, которые давали бы нам не коэффициенты c_k , но некоторые другие коэффициенты практически эквивалентной эффективности.

12. τ-метод

К счастью, по крайней мере для одного, хотя и ограниченного, но часто встречающегося в практике класса функций может быть найдено удовлетворительное решение этой задачи. Рассмотрим функцию, определенную линейным дифференциальным уравнением (или даже чисто алгебраическим уравнением), коэффициенты которого суть *рациональные функции от x* . Дифференциальное уравнение Бесселя

$$y'' + \frac{y'}{x} + \left(1 - \frac{p^2}{x^2}\right)y = 0 \quad (7-12.1)$$

и многие другие дифференциальные уравнения математической физики приводят к примерам функций этого класса. В действительности большинство важных функций, встречающихся в высших разделах физики и техники, принадлежат к этой категории ¹⁾, если добавить те функции, определение которых хотя и не дано в такой форме, но которые все же могут быть рассмотрены как решение такого дифференциального уравнения. Например, функция

$$y = \sqrt[3]{1-x} \quad (7-12.2)$$

задана так, что непосредственно не видно, что она принадлежит к этой категории, но на самом деле оказывается принадлежащей к ней; нужно только взять логарифмическую производную

$$\frac{y'}{y} = -\frac{1}{3(1-x)}, \quad (7-12.3)$$

записать это уравнение в форме

$$3(1-x)y' + y = 0 \quad (7-12.4)$$

и ввести начальное условие

$$y(0) = 1, \quad (7-12.5)$$

которое делает решение однозначным. Дифференциальное уравнение Бесселя также может быть записано без знаменателя, если умножить его на x^2 :

$$x^2 y'' + xy' + (x^2 - p^2)y = 0. \quad (7-12.6)$$

Вообще мы будем предполагать, что наше уравнение уже записано в такой форме. Например, алгебраическое определение функции (11.1) может быть заменено уравнением

$$(1+x)y - 1 = 0. \quad (7-12.7)$$

Все эти уравнения имеют некоторые общие свойства. Мы можем формально подставить степенной ряд

$$y = b_0 + b_1 x + b_2 x^2 + \dots \quad (7-12.8)$$

в такое уравнение и получить *рекуррентные соотношения* для коэффициентов. Если заданы соответствующие начальные условия, то эти коэффициенты определяются однозначно и легко могут быть вычислены. Разложение (8) почти всегда является *бесконечным* рядом. Этот ряд может сходиться и не сходиться. Если он сходится, он может сходиться в бесконечном или ограниченном интервале. Но возможно также — как это имеет место в случае так называемых «асимптотических разложений», — что формальный ряд (8) существует, но расходится для всякой точки кроме $x=0$.

¹⁾ То есть являются решениями линейных дифференциальных уравнений, коэффициенты которых суть рациональные функции x .

Теперь мы станем на другую точку зрения при изучении этих разложений, *обрывая* ряд (8) на конечном числе членов. Это значит, что мы пытаемся удовлетворить нашему уравнению полиномом

$$y = b_0 + b_1x + b_2x^2 + \dots + b_nx^n, \quad (7-12.9)$$

где n задано. Это, несомненно, может оказаться невыполнимым, так как рекуррентное соотношение, содержащее определенный, не равный нулю коэффициент b_{n+1} , не будет удовлетворяться. Можно сказать, что мы получаем «переопределенную» систему уравнений, ибо у нас $n+1$ неизвестных коэффициентов, а число уравнений, которым они должны удовлетворить, больше чем $n+1$. В случае *однородного* уравнения мы заранее знаем, что общий множитель всех коэффициентов должен остаться неопределенным. Следовательно, один из коэффициентов, например b_0 , можно принять равным 1. Число коэффициентов, которыми мы можем распорядиться, тогда не превышает n , но число уравнений, которым они должны удовлетворять, будет больше чем n .

Степень переопределенности легко можно заранее установить простым подсчетом. Например, уравнение (4) при подстановке полинома (9) не увеличивает степени старшего члена, и число рекуррентных равенств будет поэтому $n+1$. Это превышает допустимое число уравнений только *на единицу*. В случае уравнения Бесселя (6) его последний член повышает n -ю степень до $n+2$, и, таким образом, число уравнений для коэффициентов окажется равным $n+3$. Это превышает допустимое число уравнений *на три*. В случае уравнения (7) n -я степень повышается до $n+1$; следовательно, число уравнений для коэффициентов будет равно $n+2$. Здесь мы имеем *неоднородное* уравнение и можем распорядиться $n+1$ коэффициентами. Следовательно, степень переопределенности равна лишь *единице*.

Таким образом, во всех этих случаях степень переопределенности незначительна, ибо составляет небольшое *постоянное* число, тогда как число коэффициентов, подлежащих определению, может быть произвольно большим.

Эту переопределенность мы устраним теперь, введя в правую часть дифференциального уравнения некоторые члены. Это значит, что мы примиряемся с тем фактом, что не можем решить данное уравнение точно. Правую часть можно рассматривать тогда как *поправочный член*, который мы намеренно вводим для того, чтобы сделать наше уравнение разрешимым с помощью конечного полинома.

Нашей первой мыслью будет ввести в наше уравнение разложение Тейлора (8), сохранив в нем только $n+1$ членов. Так как коэффициенты определяются последовательно из рекуррентных формул, то неудовлетворенными остаются лишь *последние* из этих уравнений. Следовательно, поправочный член, который надо ввести в правую часть, будет вида τx^n в случае уравнения (4) или τx^{n+1} в случае уравнения (7), где τ есть некоторая еще неопределенная константа

В случае уравнения Бесселя (6) было бы необходимо ввести в его правую часть *три* члена вида

$$\tau_1 x^n + \tau_2 x^{n+1} + \tau_3 x^{n+2}.$$

В действительности надлежащее изучение этого дифференциального уравнения показывает, что для больших значений x мы получаем в качестве хорошего приближения

$$y = \frac{e^{ix}}{\sqrt{x}}.$$

Если мы теперь примем это асимптотическое решение за множитель при функции, дающей точное решение, т. е. примем, что

$$y = \frac{e^{ix}}{\sqrt{x}} u(x), \quad (7-12.10)$$

то тем самым определим новую функцию $u(x)$, которая гораздо более гладкая, чем была первоначальная функция $y(x)$. Для этой функции $u(x)$ определяющим дифференциальным уравнением будет

$$x^2 u'' + 2ix^2 u' - \left(p^2 - \frac{1}{4}\right) u = 0. \quad (7-12.11)$$

Это уравнение требует добавления только двух τ -членов в правую часть, ибо степень x^{n+2} исчезла. Однако получается еще и дальнейшее упрощение, если ввести новую переменную

$$\xi = \frac{1}{x} \quad (7-12.12)$$

и ввести разложение по степеням ξ , т. е. по обратным степеням x . Новое дифференциальное уравнение относительно переменной ξ имеет вид

$$\xi^2 u'' + 2(\xi - i) u' - \left(p^2 - \frac{1}{4}\right) u = 0. \quad (7-12.13)$$

Это уравнение не выводит нас за n -ю степень, и для устранения переопределенности нужен лишь один член τx^n .

Если нас интересуют *малые* значения x , можно внести аналогичное упрощение в уравнение Бесселя несколькими иными средствами. Исследование разложения Тейлора для решения обнаруживает, что в нем можно вынести за скобки общий множитель x^p . Кроме того, ряд — сомножитель содержит лишь четные степени x . Поэтому будет целесообразно ввести новую переменную

$$t = \left(\frac{x}{2}\right)^2 \quad (7-12.14)$$

и вынести общий множитель. Мы, следовательно, полагаем

$$y(2\sqrt{t}) = t^{\frac{p}{2}} u(t). \quad (7-12.15)$$

Эта подстановка дает для $u(t)$ следующее дифференциальное уравнение:

$$tu'' + (1 + p)u' + u = 0. \quad (7-12.16)$$

Снова мы отмечаем, что дифференциальное уравнение в его новой форме не вызывает повышения степени x . Поэтому добавление *одного члена* вида τx^n устранил переопределенность.

Этот пример показывает, что надлежащими преобразованиями мы можем значительно упростить нашу задачу и снизить степень переопределенности. На практике добавление *одного* τ -члена очень часто бывает достаточным, а больше *двух* τ -членов почти никогда не требуется.

Дальнейшее исследование нашей задачи обнаруживает, что нет нужды выбирать τ -член в форме τx^n . Мы можем также устранить переопределенность, выбрав правую часть в виде

$$\rho(x) = \tau p_n(x), \quad (7-12.17)$$

где $p_n(x)$ может быть произвольным полиномом. Это дает нам мощное средство повысить точность усеченного разложения Тейлора. Поправочный член

$$\rho(x) = \tau x^n, \quad (7-12.18)$$

который привел бы к усеченному разложению Тейлора, обладает тем свойством, что он остается исключительно малым в окрестности $x=0$, но затем весьма быстро возрастает в окрестности $x=1$. Следовательно, мы удовлетворили бы заданному дифференциальному уравнению с чрезвычайной точностью в окрестности начала $x=0$, поплатившись, однако, тем, что погрешность возрастает фактически как показательная функция при приближении к точке $x=1$.

Условимся, что мы ввели уже масштабный множитель, который перевел интересующий нас интервал в $[0, 1]$. Мы, несомненно, поступим куда лучше, если заменим поправочный член (18) членом

$$\rho(x) = \tau T_n^*(x), \quad (7-12.19)$$

так как полином Чебышева $T_n^*(x)$ *колеблется* в интервале $[0, 1]$ с одинаковыми амплитудами. Следовательно, мы приведем нашу погрешность на всем интервале к уравновешенной величине, которая ни слишком мала, ни слишком велика на всем интервале. Ни в какой окрестности мы не приходим к особой точности. Однако ни в какой окрестности мы не получаем и слишком большой ошибки, которая сделала бы наш ряд расходящимся. Получаемая точность эквивалентна точности, полученной с помощью «идеальных» коэффициентов Фурье (7.6), но новые коэффициенты вычисляются по *простым алгебраическим рекуррентным формулам*, без какого-либо интегрирования.

Коэффициенты этого разложения не будут уже, конечно, «жесткими» коэффициентами ряда Тейлора, образующего единый ряд чисел. Изменяя n — степень аппроксимирующего полинома — мы меняем полностью также коэффициенты полинома $T_n^*(x)$, ибо $T_n^*(x)$ заменяется теперь полиномом $T_{n+1}^*(x)$. А это означает, что, решая каждый раз другую систему рекуррентных уравнений, мы получаем совершенно новую систему коэффициентов для каждого n (ср. § 5). Наши коэффициенты соответствуют нужному нам интервалу, а *также* степени того полинома, с которым мы оперируем. Результат этой «гибкости» тройкий:

1. Мы можем улучшить сходимость разложения Тейлора, т. е. мы можем сделать погрешность нашего приближения гораздо меньшей, чем погрешность разложения Тейлора, при том же числе членов.

2. Мы можем получить сходящееся разложение в случаях, когда разложение Тейлора *расходится* либо в заданном интервале, либо в его начале, либо начиная с какой-нибудь его точки. Все так называемые «асимптотические» разложения обычных трансцендентных функций математической физики могут быть таким образом преобразованы из расходящихся в сходящиеся разложения.

3. Мы можем получить сходящееся разложение в случаях, когда ряда Тейлора совсем не существует.

13. Канонические полиномы

Против оперирования с «гибкими» коэффициентами можно выдвинуть следующее возражение. Предположим, что нас не удовлетворило приближение некоторой определенной степени. Тогда при переходе от n к $n+1$ нам придется отбросить полностью все предыдущие результаты и снова начать все вычисления. Мы сейчас изложим метод, который свободен от этого недостатка.

Вместо того чтобы добавлять в правую часть поправочный член в форме $\tau T_n^*(x)$, мы добавим его в правую часть в виде одной степенной функции x^m и решим получающиеся уравнения для коэффициентов. Тогда, произведя подсчет числа уравнений, подлежащих решению, мы обнаружим, что число наложенных условий превышает число параметров, которыми мы можем распорядиться, на некоторое число μ . Мы устраним это переопределение, *опустив* на время *первые* μ *уравнений*. Мы рассматриваем их как своего рода «начальные условия», о которых позаботимся позднее.

Для иллюстрации исследуем дифференциальное уравнение

$$x^3 y' - 2y = 0. \quad (7-13.1)$$

К этому уравнению мы пришли бы, если бы хотели получить степенной ряд для функции

$$y = e^{-\frac{1}{x^2}}. \quad (7-13.2)$$

Рекуррентные соотношения между коэффициентами приводят к $n+3$ уравнениям для $n+1$ коэффициентов. Избыточных уравнений, таким образом, *два*. Поэтому мы опускаем два начальных уравнения, для того чтобы сделать систему совместной. Если подставим в наше уравнение формальное разложение (12.8) и образуем таблицу результирующих коэффициентов, то получим последовательно

$$\begin{aligned} -2b_0 &= 0, \\ \boxed{\begin{aligned} -2b_1 &= 0, \\ -2b_2 &= 0, \end{aligned}} \\ b_1 - 2b_3 &= 0, \\ 2b_2 - 2b_4 &= 0, \\ 3b_3 - 2b_5 &= 0, \\ 4b_4 - 2b_6 &= 0, \\ \dots \end{aligned} \quad (17-13.3)$$

От переопределения мы освобождаемся, опустив два уравнения, взятых в рамку.

Но поправочный член x^m в правой части уравнения (1) означает появление «единицы» в правой части одного из уравнений (3). Эта единица появится в m -м уравнении. Давая m последовательно значения $0, 1, 2, \dots$, мы заставим эту единицу скользить сверху вниз. Так как, однако, в нашей задаче мы вычеркнули второе и третье уравнения, то значения $m=1$ и $m=2$ мы пропускаем.

Таким образом, мы получаем определенный ряд полиномов, внутренне связанных с заданным дифференциальным оператором. Мы называем их «каноническими полиномами» и обозначаем через $Q_m(x)$. Эти $Q_m(x)$ не обязательно будут иметь m -ю степень. Индекс m отмечает просто тот факт, что если выполнить операции, указанные в левой части заданного дифференциального уравнения, над полиномом $Q_m(x)$ (опустив лишние уравнения), то в результате получим x^m *). В нашей задаче, например, мы воздержимся от опреде-

*) Несколько подробнее и точнее: при подстановке в наше дифференциальное уравнение вместо y полинома $Q_m(x)$ будут удовлетворяться все отобранные уравнения (3), за исключением опущенных, причем в последнем из отобранных уравнений в правой части вместо 0 стоит 1. Соответственно этому в правой части дифференциального уравнения будет член x^m и, возможно, еще члены, соответствующие коэффициентам, которые входят в опущенные уравнения (3). Например, пусть в нашей задаче мы выбираем $m=5$. Мы, следовательно, принимаем за поправочный член x^5 . Отбираем первые шесть уравнений для коэффициентов из системы (3). Однако полином $Q_5(x)$, чтобы удовлетворить видоизмененному уравнению (1) с таким поправочным членом, должен иметь степень 3. Мы можем распорядиться, следовательно, четырьмя коэффициентами. Поэтому из шести уравнений для коэффициентов мы опус-

ления $Q_1(x)$ и $Q_2(x)$, но остальные $Q_k(x)$ однозначно определяются:

$$\begin{aligned} Q_0(x) &= -\frac{1}{2}, \\ Q_3(x) &= x, \\ Q_4(x) &= \frac{1}{2}x^2, \\ Q_5(x) &= \frac{1}{3}(2x + x^3), \\ Q_6(x) &= \frac{1}{4}(x^2 + x^4), \\ Q_7(x) &= \frac{1}{15}(4x + 2x^3 + 3x^5), \\ &\dots \end{aligned} \tag{7-13.4}$$

Другой интересный пример представляет дифференциальное уравнение

$$xy' - y - x = 0, \tag{7-13.5}$$

которое определяет функцию

$$y = x \lg x, \tag{7-13.6}$$

если не считать аддитивного члена cx . Здесь простой подсчет числа уравнений показывает, что переопределения нет, так как получается $n+1$ уравнений для $n+1$ коэффициентов, и система не однородна. Но второе уравнение для коэффициентов дает противоречивое равенство $-1=0$. Мы рассматриваем его как переопределение и опускаем из нашей системы.

В этом примере каноническими полиномами будут

$$\begin{aligned} Q_0(x) &= -1, \\ Q_2(x) &= x^2, \\ Q_3(x) &= \frac{x^3}{2}, \\ &\dots \\ Q_m(x) &= \frac{x^m}{m-1}. \end{aligned} \tag{7-13.7}$$

каем два (окаймленные). Последним отобранным уравнением будет тогда $3b_3 - 2b_5 = 1$. Так как степень $Q_5(x)$ есть 3, то $b_5 = b_4 = 0$. Для остальных коэффициентов получаем из отобранных уравнений $b_3 = \frac{1}{3}$; $b_1 = \frac{2}{3}$; $b_2 = b_0 = 0$. Это и определяет $Q_5(x)$, как указано в формулах (4).

Подставив этот полином $Q_5(x)$ в уравнение (1), получим в правой части $x^5 - \frac{4}{3}x$. Здесь, кроме нужного члена x^5 , получен еще член $-\frac{4}{3}x$ соответственно тому факту, что второе из уравнений (3) $-2b_1 = 0$ не выполнено. При других значениях m могло бы неудовлетворяться и условие $b_2 = 0$. Эти два дополнительных условия будут удовлетворены надлежащим выбором неопределенных коэффициентов τ_1 и τ_2 , как указано далее в тексте.

Во всех случаях мы можем без труда построить шаг за шагом канонические полиномы, даже если не всегда можно записать их общее выражение столь просто, как в последнем примере.

Эти канонические полиномы дают нам возможность получить решение дифференциального уравнения $Dy = 0$, если добавить в правую часть его поправочный член вида (12.19). Мы используем теперь принцип суперпозиции линейных операторов. Так как полином $T_n^*(x)$ может быть выписан со своими коэффициентами

$$T_n^*(x) = \sum_{m=0}^n c_n^m x^m, \quad (7-13.8)$$

то решение уравнения

$$Dy = \tau T_n^*(x) \quad (7-13.9)$$

получает вид

$$y = \tau \sum_{m=0}^n c_n^m Q_m(x). \quad (7-13.10)$$

Решение в явной форме оказывается линейной суперпозицией *неизменного ряда полиномов*, а именно канонических полиномов, однозначно связанных с заданным дифференциальным оператором. Коэффициенты полиномов Чебышева входят просто в качестве *весов* этих полиномов. Постепенное решение уравнений для коэффициентов отдельно для каждого значения n теперь уже не нужно. Решение (10) можно выразить в следующей условной форме:

$$y = \tau T_n^*(Q(x)), \quad (7-13.11)$$

причем здесь в формальном разложении полинома $T_n^*(x)$ m -я степень x заменяется через $Q_m(x)$.

Свобода выбора параметра τ может быть теперь использована, чтобы удовлетворить излишнему уравнению, ранее опущенному. Если пришлось опустить больше одного уравнения, то необходимо добавить в правую часть более одного τ -члена. Например, в случае *двух* избыточных уравнений понадобятся поправочные члены:

$$\rho(x) = \tau_1 T_n^*(x) + \tau_2 T_{n+1}^*(x),$$

которые содержат два τ -множителя, подлежащие определению из двух оставшихся условий. На практике удобнее оперировать только с одним полиномом Чебышева. Этого можно достичь, пользуясь поправочным членом следующего вида:

$$\rho(x) = T_n^*(x) (\tau_1 + \tau_2 x). \quad (7-13.12)$$

Явная форма решения (13.10) принимает теперь вид

$$y_n = \sum_{m=0}^n c_n^m [\tau_1 Q_m(x) + \tau_2 Q_{m+1}(x)]. \quad (7-13.13)$$

Оставшиеся два уравнения дают систему линейных уравнений относительно τ_1 и τ_2 , которые без труда могут быть решены.

Мы проведем все решение для случая дифференциального уравнения (5). Здесь решение получается в виде

$$y_n = -\tau + \tau \sum_{m=2}^n c_n^m \frac{x^m}{m-1} + cx, \quad (7-13.14)$$

где c — произвольная постоянная. Уравнение, которое было опущено, есть

$$-1 = \tau c_n^1, \quad (7-13.15)$$

откуда

$$\tau = -\frac{1}{c_n^1} = \frac{(-1)^n}{2n^2}. \quad (7-13.16)$$

Постоянная c определяется из граничного условия

$$y(1) = 0. \quad (7-13.17)$$

Это решение дает улучшающие приближения функции $y = x \log x$ в форме степенного разложения с гибкими (изменяющимися) коэффициентами. Для функции $x \log x$ не существует ряда Тейлора ввиду аналитической особенности в точке $x = 0$. Между тем разложение (14) сходится к $y = x \log x$ равномерно во всем интервале $[0, 1]$. Более того, сходимость удовлетворительна даже при малых n , ибо максимальная ошибка ограничена величиной $|\tau|$, т. е. $\frac{1}{2n^2}$. Например, при выборе $n = 4$ получаем разложение

$$\begin{aligned} y_4 &= -\frac{1}{32} + \frac{1}{32} \left(160x^2 - \frac{256}{2}x^3 + \frac{128}{3}x^4 \right) - \frac{221}{96}x = \\ &= \frac{1}{96} \left(-32 - 221x + 480x^2 - 384x^3 + 128x^4 \right) \end{aligned} \quad (7-13.18)$$

с максимальной погрешностью $\pm 0,03$.

Пользование каноническими полиномами $Q_n(x)$ можно рекомендовать еще и с другой точки зрения. Оно дает возможность приспособить решение для произвольных интервалов. Допустим, что нам нужно решение в интервале $[0, \alpha]$ вместо $[0, 1]$. Это α может быть любым вещественным или комплексным числом, если только луч $[0, \alpha]$ не содержит особой точки функции $y(x)$. Приспособление интервала выполняется сразу же простой заменой $T_n^*(x)$ полиномом $T_n^*\left(\frac{x}{\alpha}\right)$.

Это означает, что коэффициенты c_n^m должны быть заменены числами $c_n^m \alpha^{-m}$. Следовательно, решение (10) примет теперь форму

$$y(x) = \tau \sum_{m=0}^n c_n^m \alpha^{-m} Q_m(x), \quad (7-13.19)$$

где x есть любая точка на комплексном луче $[0, \alpha]$. Если мы

дойдем до конечной точки интервала $x = \alpha$, то получим решение заданного дифференциального уравнения в произвольной точке z комплексной плоскости, совместимое с общими условиями регулярности. Решение получится в виде

$$y(z) = \tau \sum_{m=0}^n c_n^m z^{-m} Q_m(z) = \tau T_n^* \left(\frac{Q(z)}{z} \right). \quad (7-13.20)$$

Таким образом, мы обошли ограничение, что x должен лежать между 0 и 1. Однако это решение вообще не будет уже полиномом, а отношением двух полиномов (ср. § 14, пример 5).

14. Примеры приложения τ -метода

τ -метод имеет столь широкое поле приложений, что полезно разобрать несколько характерных примеров, показывающих его силу. Проблему определения соответствующих границ погрешности мы рассмотрим в следующем параграфе.

Первый пример имеет дело с функцией, для которой разложение Тейлора хорошо сходится. Мы покажем, как применение τ -метода намного улучшает сходимость. Во втором примере мы покажем, как область сходимости ряда Тейлора может быть распространена на более обширный интервал. Третий пример покажет, как полностью расходящееся «асимптотическое» разложение может быть заменено сходящимся разложением. В четвертом примере будет рассмотрена функция, для которой не существует разложения Тейлора. Пятый пример проиллюстрирует распространение метода на комплексную область.

Пример 1. Рассмотрим функцию

$$y = e^x, \quad (7-14.1)$$

определяемую дифференциальным уравнением

$$y - y' = 0, \quad (7-14.2)$$

с граничным условием

$$y(0) = 1. \quad (7-14.3)$$

Мы имеем $n+2$ уравнения с $n+1$ коэффициентами. Уравнения становятся совместными, если просто опустить граничное условие.

Каноническими полиномами $Q_m(x)$ нашей задачи будут

$$Q_0(x) = 1,$$

$$Q_1(x) = 1 + x,$$

$$Q_2(x) = 2 \left(1 + x + \frac{x^2}{2} \right),$$

$$Q_3(x) = 3! \left(1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} \right), \quad (7-14.4)$$

.....

$$Q_m(x) = m! \left(1 + x + \frac{x^2}{2!} + \dots + \frac{x^m}{m!} \right) = m! S_m(x),$$

где

$$S_m(x) = 1 + x + \frac{x^2}{2!} + \dots + \frac{x^m}{m!}$$

представляют собой последовательные «частные суммы» ряда Тейлора. Следовательно, решение нашей задачи согласно (13.10) будет

$$y_n(x) = \tau \sum_{m=0}^n c_n^m m! S_m(x). \quad (7-14.5)$$

Теперь мы удовлетворяем условию (3) и получаем

$$y_n = \frac{\sum_{m=0}^n c_n^m m! S_m(x)}{\sum_{m=0}^n c_n^m m!}. \quad (7-14.6)$$

Решение оказывается *взвешенным арифметическим средним частных сумм ряда Тейлора*. Последовательные приближения функций e^x в интервале $[0,1]$, таким образом, будут

$$\begin{aligned} y_0 &= 1, \\ y_1 &= \frac{1 + 2x}{1}, \\ y_2 &= \frac{9 + 8x + 8x^2}{9}, \\ y_3 &= \frac{113 + 114x + 48x^2 + 32x^3}{113}, \\ y_4 &= \frac{1825 + 1824x + 928x^2 + 256x^3 + 128x^4}{1825}. \end{aligned} \quad (7-14.7)$$

Соответствующие значения для e получаются, если положить $x=1$:

$$1; 3; \quad 2,777; \quad 2,7168; \quad 2,71836; \quad 2,71828068; \dots \quad (7-14.8)$$

(точка под цифрой указывает первый десятичный знак погрешности). Эта последовательность сходится гораздо быстрее последовательности частных сумм ряда Тейлора:

$$1; \quad 2; \quad 2,5; \quad 2,667; \quad 2,70867; \quad 2,71667; \dots,$$

и все же эти результаты не могут сравниться с эффективными приближениями числа e , найденными в гл. VI, 20. Прежние приближения аналогичны найденным здесь и могут быть получены с помощью той же процедуры, если только коэффициенты c_n^m полиномов Чебышева заменить соответствующими коэффициентами полиномов Лежандра $P_n^*(x)$. Так как мы установили, что полиномы Чебышева дают меньшую погрешность, чем полиномы Лежандра, то кажется странным, что прежние значения величины e настолько превосходят значения,

полученные теперь. Однако следует иметь в виду *весь* интервал x между 0 и 1. Малость погрешности в конечной точке $x=1$ не гарантирует, что погрешность будет малой *всюду*. Кроме того, выбор $T_n^*(x)$ в качестве поправочного члена дифференциального уравнения, обеспечивая хорошую сходимость, не обязательно дает *наилучшую* возможную сходимость. Мы добивались равномерного колебания погрешности приближения *функции* $y=f(x)$. Операция Dy может значительно изменить этот равномерный характер погрешности. При надлежащем изучении данного дифференциального оператора мы можем составить такой поправочный член $\rho(x)$, который окажется лучшим, чем выбранный нами член $\tau T_n^*(x)$.

Например, в нашей задаче можно без труда заметить, что доминирующим членом в заданном дифференциальном уравнении является y' , так как дифференцирование высокочастотного колебания типа $\cos n\theta$ увеличивает амплитуду в n раз. Поэтому в этой задаче предпочтительнее было бы положить

$$\rho(x) = \tau T_{n+1}^{*'}(x), \quad (7-14.9)$$

ибо если принять, что многочленное приближение $y_n(x)$ имеет форму

$$y_n(x) = e^x - \tau T_{n+1}^{*'}(x), \quad (7-14.10)$$

то операция $y - y'$ образует (с хорошим приближением) поправочный член (9). Решение (5) будет действительным опять, но коэффициенты c_n^m должны быть заменены коэффициентами c_n^m полиномов $T_{n+1}^{*'}(x)$, называемых «полиномами Чебышева второго рода». Кроме того, предполагаемое решение (10) указывает, что мы не удовлетворим в точности граничному условию $y(0)=1$, а определим τ из условия

$$y_n(0) = 1 - \tau T_{n+1}^{*'}(0) = 1 - \tau (-1)^{n+1}, \quad (7-14.11)$$

которое дает:

$$\tau [(-1)^{n+1} + \sum_{m=0}^n \bar{c}_n^m m!] = 1. \quad (7-14.12)$$

Таким образом, окончательное решение будет иметь вид

$$y_n(x) = \frac{\sum_{m=0}^n \bar{c}_n^m m! S_m(x)}{(-1)^{n+1} + \sum_{m=0}^n \bar{c}_n^m m!}. \quad (7-14.13)$$

Это опять взвешенное арифметическое среднее частных сумм ряда Тейлора, но веса теперь другие. Первые пять приближений для e^x

в интервале $[0,1]$ теперь получаются следующие:

$$\begin{aligned} y_0 &= \frac{2}{1}, \\ y_1 &= \frac{8 + 16x}{9}, \\ y_2 &= \frac{114 + 96x + 96x^2}{113}, \\ y_3 &= \frac{1824 + 1856x + 768x^2 + 512x^3}{1825}, \\ y_4 &= \frac{36690 + 36640x + 18720x^2 + 5120x^3 + 2560x^4}{36689}. \end{aligned} \quad (7-14.14)$$

Ориентировочная максимальная погрешность в любой точке интервала не превышает числа, обратного знаменателю. Следовательно, квадратичное приближение ограничивается $\pm 0,01$, а кубическое — лучше чем $\pm 0,0006$. Первые 6 значений e получаются такие:

$$2; \quad 2,667; \quad 2,70796; \quad 2,717808; \quad 2,718253; \quad 2,71828075.$$

«Наилучший» характер этого приближения остается действительным, несмотря на превосходство замечательных приближений (6-20.4) и (6-20.5), полученных на основе квадратичного метода, изложенного в гл. V.22. То были приближения в конечной точке, и можно показать, что для нашей задачи полиномы Лежандра превосходят полиномы Чебышева с точки зрения минимизирования погрешности для $y(1)$ в конечной точке интервала.

Пример 2. Рассмотрим функцию

$$y = \frac{1}{a + x}, \quad (7-14.15)$$

определяемую алгебраическим уравнением

$$(a + x)y - 1 = 0. \quad (7-14.16)$$

Вместо использования канонических полиномов, мы в этой задаче будем следовать другим путем, чтобы проиллюстрировать метод, который в некоторых случаях оказывается более полезным. Вместо разложения $y(x)$ по степеням x мы сразу располагаем наше разложение по полиномам Чебышева:

$$y = \frac{1}{2} c_0 T_0^*(x) + c_1 T_1^*(x) + \dots + c_{n-1} T_{n-1}^*(x). \quad (7-14.17)$$

Как обычно, мы для обеспечения совместности добавляем в правой части член с τ

$$(a + x)y = 1 + \tau T_n^*(x). \quad (7-14.18)$$

В этой задаче мы можем сразу определить τ . Подставляя $x = -a$, получаем в левой части нуль. Следовательно,

$$\tau = -\frac{1}{T_n^*(a)}. \quad (7-14.19)$$

Теперь мы подставляем разложение (17) в левую часть и выполняем умножение на $(a \mp x)$. Воспользуемся рекуррентным соотношением (7.7), которое для смещенных полиномов $T_n^*(x)$ имеет вид

$$2(2x - 1) T_k^*(x) = T_{k+1}^*(x) \mp T_{k-1}^*(x). \quad (7-14.20)$$

Умножим уравнение (18) на 4, полагая

$$4a \mp 2 = 2b. \quad (7-14.21)$$

Тогда

$$[2b \mp (4x - 2)]y = 4 \mp 4\tau T_n^*(x). \quad (7-14.22)$$

Образую таблицу коэффициентов разложения, получаем последовательно:

в левой части

	bc_0	$2bc_1$	$2bc_2$	$2bc_3$	
		c_0	c_1	c_2	
в правой части	$\frac{c_1}{4}$	$\frac{c_2}{0}$	$\frac{c_3}{0}$	$\frac{c_4}{0}$	$\dots$

Это дает уравнения

$$\begin{aligned} bc_0 \mp c_1 &= 4, \\ c_0 \mp 2bc_1 \mp c_2 &= 0, \\ c_1 \mp 2bc_2 \mp c_3 &= 0, \\ &\dots \end{aligned} \quad (7-14.23)$$

Оставим на время в стороне первое уравнение. Остальные однородные уравнения можно решить, если положить, что

$$c_k = Cp^k. \quad (7-14.24)$$

Тогда все уравнения приводятся к одному квадратному уравнению относительно p

$$1 \mp 2bp \mp p^2 = 0, \quad (7-14.25)$$

два корня которого суть

$$p = -b \pm \sqrt{b^2 - 1} = -(2a \mp 1) \pm 2\sqrt{a(a \mp 1)}. \quad (7-14.26)$$

Соответственно принципу суперпозиции для решения линейной системы мы получаем, таким образом,

$$c_k = C_1 p_1^k + C_2 p_2^k, \quad (7-14.27)$$

где p_1 и p_2 суть два корня, соответствующие знаку $\pm$ в формуле (26). Две константы C_1 и C_2 пока произвольны. Но условие $c_n = 0$ дает

$$C_1 p_1^n + C_2 p_2^n = 0, \quad (7-14.28)$$

между тем как первое уравнение системы (23) требует

$$C_1 p_1 + C_2 p_2 + b(C_1 + C_2) = 4. \quad (7-14.29)$$

Квадратное уравнение (25) показывает, что произведение двух корней p_1 и p_2 равно 1. Обозначим через p_1 меньший, а через p_2 — больший по абсолютной величине корень. Тогда из соотношения (28) получаем

$$C_2 = -C_1 \left(\frac{p_1}{p_2} \right)^n. \quad (7-14.30)$$

Когда n возрастает, C_2 стремится к нулю. Мы упростим положение, не увеличивая существенно ошибки, если совсем опустим C_2 . Это означает, что мы сразу устанавливаем *бесконечное* разложение по $T_n^*(x)$, т. е. ряд Фурье, если мыслить это разложение составленным относительно переменной θ [ср. (8.3)], и затем обрываем этот ряд после n членов. Наше решение тогда будет записываться просто:

$$c_k = \frac{4}{b+p} p^k = \frac{2}{\sqrt{a(a+1)}} p^k, \quad (7-14.31)$$

где

$$p = 2\sqrt{a(a+1)} - (2a+1) \quad (7-14.32)$$

в предположении, что a — положительное число.

Вообще, a может быть любым вещественным или комплексным постоянным. Совершенно общим решением нашей задачи будет бесконечное разложение

$$y = \frac{1}{a+x} = \frac{2}{\sqrt{a(a+1)}} \left[\frac{1}{2} T_0^*(x) + p T_1^*(x) + p^2 T_2^*(x) + \dots \right], \quad (7-14.33)$$

где p — меньший по абсолютной величине корень из (26). Если этот ряд оборвать после n членов, то остаток ряда можно оценить на основании того факта, что все $T_k^*(x)$ колеблются между ± 1 :

$$|\eta_n| \leq \frac{2|p|^n}{\sqrt{a(a+1)}} (1 + |p| + |p|^2 + \dots),$$

$$|\eta_n| \leq \frac{2|p|^n}{\sqrt{a(a+1)}} \frac{1}{1-|p|}. \quad (7-14.34)$$

Так как $p_1 p_2 = 1$, то один из корней всегда остается по абсолютной величине меньше единицы, за исключением предельного случая, когда p_1 и p_2 оба лежат на единичном круге и являются комплексно сопряженными величинами. Это возможно лишь, если a есть *вещественное отрицательное число*, содержащееся между 0 и -1 . В этом случае $y(x)$ имеет особую точку в критическом интервале $[0,1]$. Во всех остальных случаях ряд (33) сходится даже абсолютно.

Сравним этот результат с рядом Тейлора, который сходится только до $x = |a|$ и расходится при $|x| > |a|$. В обоих случаях разложение носит характер геометрической прогрессии и оценка максимальной погрешности η_n в любой точке интервала имеет вид (34). Но p в этой формуле имеет в обоих случаях весьма различные значения. Нижеследующая таблица сопоставляет численные значения p в этих двух случаях для широкого интервала параметра a при условии, что a — вещественное положительное число:

a	Ряд Тейлора — p	$r_n^*(x)$ — p
3	0,333	0,072
2	0,5	0,101
1	1	0,172
0,5	2	0,263
0,333	3	0,333
0,2	5	0,420
0,1	10	0,537
0,01	100	0,819

(7-14.35)

Особенно интересен случай

$$\begin{aligned}
 y &= \frac{1}{1+x} = 1 - x + x^2 - x^3 + \dots = \\
 &= \sqrt{2} \left[\frac{1}{2} - 0,1716 T_1^*(x) + (0,1716)^2 T_2^*(x) - \right. \\
 &\quad \left. - (0,1716)^3 T_3^*(x) + \dots \right].
 \end{aligned}
 \tag{7-14.36}$$

Знаменатель второго разложения (прогрессии) получен из равенства

$$3 - 2\sqrt{2} = 0,1716 \dots$$

Сравнительно с рядом Тейлора мы уменьшили его в 5,83 раза. Грубая общая оценка¹⁾ дает уменьшение в 4 раза и это фактически верно для больших значений a , для которых ряд Тейлора еще

¹⁾ См. [4].

хорошо сходится, как можно показать на примере $a=3$. При убывании a выигрыш медленно увеличивается и достигает на пределе сходимости ряда Тейлора максимального значения

$$3 + 2\sqrt{2} = 5,8284 \dots$$

При $a=1$ ряд Тейлора расходится во всем интервале $[0,1]$, тогда как разложение по полиномам $T_n^*(x)$ сохраняет свою сходимость.

Пример 3. Этот пример выбран потому, что он показывает, каким образом τ -метод преобразовывает совершенно расходящееся разложение в строго сходящееся. Рассмотрим «интегральную показательную функцию»

$$\omega(z) = \int_z^\infty \frac{e^{-t}}{t} dt, \quad (7-14.37)$$

встречавшуюся нам уже в (3.1). Как и раньше, полагаем

$$\omega(z) = \frac{e^{-z}}{z} u(z) \quad (7-14.38)$$

и получаем для новой функции $u(z)$ дифференциальное уравнение

$$-zu' + (1+z)u = z. \quad (7-14.39)$$

Теперь вводим в качестве новой независимой переменной

$$x = \frac{1}{z} \quad (7-14.40)$$

и получаем новое дифференциальное уравнение

$$x^2 y' + (1+x)y = 1. \quad (7-14.41)$$

Если $y(x)$ формально разложить по степеням x^*), то получим следующий бесконечный ряд:

$$y(x) = 1 - x + 2!x^2 - 3!x^3 + 4!x^4 - \dots \quad (7-14.42)$$

Переменная x в этом разложении на самом деле есть обратная величина первоначального переменного z . Следовательно, можно понимать разложение (42) как формальный антипод разложения

$$e^{-z} = 1 - z + \frac{z^2}{2!} - \frac{z^3}{3!} + \dots, \quad (7-14.43)$$

поскольку каждый член первого ряда является точным обращением соответствующего члена второго разложения. В то время как ряд (43) исключительно хорошо сходится, оставаясь сходящимся для любого значения переменного z — ряд (42) совершенно расходится; сходимости нет даже при сколь угодно малом значении x , за исклю-

*) При начальном условии $y(0) = 1$.

чением тривиального значения $x=0$. Тем не менее такой ряд имеет большое значение, так как можно показать, что погрешность усеченного ряда меньше, чем первый опущенный член. Поэтому мы пользуемся только частью ряда, содержащей убывающие члены, останавливаясь на надлежащем члене. Этот член мы определяем из условия, что первый опущенный член должен быть меньше следующего за ним. По этому методу мы получаем хорошее приближение для достаточно малых значений x , но с увеличением x ряд постепенно утрачивает свое значение.

Теперь мы решим эту задачу τ -методом. В системе уравнений для коэффициентов мы опускаем первое с целью сделать систему совместной и определяем канонические полиномы $Q_m(x)$, начиная с $m=1$. Таким образом, мы находим

$$Q_m(x) = \frac{(-1)^{m-1}}{m!} S_{m-1}(x)^*, \quad (7-14.44)$$

а τ -решение принимает вид

$$y_{n-1}(x) = \tau \sum_{m=1}^n c_n^m (-1)^{m-1} \frac{S_{m-1}(x)}{m!}. \quad (7-14.45)$$

Теперь мы удовлетворяем временно опущенному первому уравнению

$$b_0 = 1 + \tau c_n^0. \quad (7-14.46)$$

Это дает

$$\tau = \frac{1}{\sum_{m=0}^n \frac{c_n^m (-1)^{m-1}}{m!}}, \quad (7-14.47)$$

а полным решением служит

$$y_{n-1}(x) = \frac{\sum_{m=1}^n (-1)^{m-1} c_n^m \frac{S_{m-1}(x)}{m!}}{\sum_{m=0}^n (-1)^{m-1} c_n^m \frac{1}{m!}}. \quad (7-14.48)$$

Интересно сопоставить это решение с решением (6), найденным в случае показательной функции. В обоих случаях решение имеет вид взвешенного арифметического среднего частных сумм ряда Тейлора. Существенное отличие состоит в том, что множитель $m!$ раньше был в числителе каждого члена, теперь же он входит в знаменатель. Раньше главный упор делался на частные суммы высших порядков, теперь эти суммы играют весьма малую роль.

*) S_m — здесь, как и во всем параграфе, суть частные суммы членов ряда Тейлора.

Пользование усеченным рядом Тейлора порядка n соответствует крайнему случаю выбора весов: $0, 0, \dots, 0, 1$. В случае хорошо сходящейся функции e^x такой выбор не вредит и не очень далек от взвешивания при более эффективном τ -ряде. В последнем веса не столь резко отличаются между собой, но все же аналогичны весам для ряда Тейлора, поскольку центр тяжести переносится на суммы, соответствующие большим значениям m . В нашем примере положение совсем иное. Здесь влияние членов высших порядков снижается малостью весов, и центр тяжести переносится на частные суммы низших порядков. Это соответствует обычному приему образования частных сумм до определенного члена, когда удовлетворяются некоторой погрешностью, которая не может быть уменьшена. Однако при τ -методе нет нужды останавливаться на определенном значении n . Частные суммы высших порядков не отбрасываются, но их вредное влияние снижается надлежащим множителем. Суммы высших порядков служат корректирующими членами, и по мере возрастания n мы действительно все ближе и ближе подходим к точному значению функции в любой точке интервала. Будут ли первоначальные суммы сходиться сами по себе или нет — безразлично. *Надлежащие веса делают их средние арифметические сходящимися во всяком случае.*

Первыми пятью приближениями в нашей задаче будут

$$y_0 = \frac{2}{1+2} = \frac{2}{3},$$

$$y_1 = \frac{8 + \frac{8}{2}(1-x)}{1 + 8 + \frac{8}{2}} = \frac{12 - 4x}{13},$$

$$y_2 = \frac{18 + \frac{48}{2}(1-x) + \frac{32}{6}(1-x+2x^2)}{1 + 18 + \frac{48}{2} + \frac{32}{6}} = \frac{142 - 88x + 32x^2}{145}, \quad (7-14.49)$$

$$y_3 = \frac{32 + \frac{160}{2}(1-x) + \frac{256}{6}(1-x+2x^2) + \frac{128}{24}(1-x+2x^2-6x^3)}{1 + 32 + \frac{160}{2} + \frac{256}{6} + \frac{128}{24}} =$$

$$= \frac{160 - 128x + 96x^2 - 32x^3}{161},$$

$$y_4 = \frac{50 + \frac{400}{2}(1-x) + \frac{1120}{6}(1-x+2x^2) + \frac{1280}{24} \left(\begin{array}{c} \\ \end{array} \right) + \frac{512}{120} \left(\begin{array}{c} \\ \end{array} \right)}{1 + 50 + \frac{400}{2} + \frac{1120}{6} + \frac{1280}{24} + \frac{512}{120}} =$$

$$= \frac{7414 - 6664x + 7328x^2 - 5184x^3 + 1536x^4}{7429}.$$

Оценка точности решения (ср. § 15) такова, что максимальная по-

грешность в любой точке интервала не может превысить τ . Поэтому

$$|\eta| \leq 0,33; \quad 0,077; \quad 0,021; \quad 0,0062; \quad 0,0021; \quad 0,00070; \dots$$

В конечной точке $x=1$ интервала мы получаем следующие приближения для $y(1)=0,596358414$:

$$0,667; \quad 0,6154; \quad 0,5931; \quad 0,59627; \quad 0,596312; \quad 0,5963666. \quad (7-14.50)$$

Асимптотический ряд указывает лишь, что значение функции в этой точке должно лежать между 0 и 1. Напротив, τ -решение дает по мере возрастания τ значение функции с произвольной точностью.

Пример 4. Теперь мы получим разложение для функции

$$y = \sqrt{x}. \quad (7-14.51)$$

Так как все производные функции становятся бесконечно большими в точке $x=0$, то в этом случае ряд Тейлора не существует.

Взяв логарифмическую производную, получаем

$$\frac{y'}{y} = \frac{1}{2x}. \quad (7-14.52)$$

Мы видим, следовательно, что нашу функцию можно определить дифференциальным уравнением

$$2xy' - y = 0, \quad (7-14.53)$$

но мы заменяем его уравнением

$$2xy' - y = \tau T_n^*(x). \quad (7-14.54)$$

Каноническими полиномами в нашей задаче являются

$$\begin{aligned} Q_0 &= -1, \\ Q_1 &= x, \\ Q_2 &= \frac{x^2}{3}, \\ &\dots \dots \dots \\ Q_m &= \frac{x^m}{2m-1} \end{aligned} \quad (7-14.55)$$

и, следовательно,

$$y_n(x) = \tau \sum_{m=0}^n \frac{c_n^m}{2m-1} x^m. \quad (7-14.56)$$

Множитель τ может быть определен из граничного условия

$$y(1) = 1. \quad (7-14.57)$$

Отсюда

$$y_n(x) = \frac{\sum_{m=0}^n \frac{c_n^m}{2m-1} x^m}{\sum_{m=0}^n \frac{c_n^m}{2m-1}}. \quad (7-14.58)$$

Исследование погрешности в этой задаче особенно поучительно. Критической точкой является точка $x=0$, где функция имеет аналитическую особенность. Поэтому наибольшие погрешности будут в окрестности начала. Мы будем оперировать с угловой переменной θ (§ 8), которая является подходящей переменной всякий раз, когда привлекаются полиномы Чебышева. В целях удобства мы заменяем θ на $\pi - \theta$ и получаем соответственно этому

$$x = \sin^2 \frac{\theta}{2},$$

$$T_n^*(x) = (-1)^n \cos n\theta.$$

Для малых x мы можем положить

$$x = \frac{\theta^2}{4} \quad (7-14.59)$$

и переписать дифференциальное уравнение (54) с переменной θ :

$$\theta y' - y = \bar{\tau} \cos n\theta, \quad (7-14.60)$$

где $\bar{\tau} = (-1)^n \tau$. Это дифференциальное уравнение мы решаем методом «варьирования произвольных постоянных». Полагаем

$$y = C\theta, \quad (7-14.61)$$

$$C' = \bar{\tau} \frac{\cos n\theta}{\theta^2}. \quad (7-14.62)$$

Интегрируя по частям, получаем

$$C = -\frac{\bar{\tau}}{\theta} \cos n\theta - n\bar{\tau} \int \frac{\sin n\theta}{\theta} d\theta. \quad (7-14.63)$$

Отсюда

$$y_n = -\bar{\tau} \cos n\theta - n\bar{\tau} \theta \left[\int_0^\theta \frac{\sin n\theta}{\theta} d\theta + A \right], \quad (7-14.64)$$

где A есть постоянная интегрирования. Так как, однако, функция $y(\theta)$ должна быть полиномом относительно θ^2 , то члена вида $A\theta$ не может быть. Это показывает, что A надо положить равным нулю, и мы получаем

$$y_n(\theta) = -\bar{\tau} \cos n\theta - \bar{\tau} n\theta \operatorname{si}(n\theta). \quad (7-14.65)$$

Функция $\text{si}(n\theta)$ колеблется с уменьшающимися амплитудами вокруг асимптотического значения $\frac{\pi}{2}$. Это показывает, что $y(\theta)$ колеблется около функции

$$-\bar{\tau}n\frac{\pi}{2}\theta = -\bar{\tau}n\pi\sqrt{x}.$$

Так как функция, которую нужно найти, есть $\sqrt{x}$, нам нужно выбрать $\bar{\tau}$ так, чтобы множитель при $\sqrt{x}$ сделался равным 1. Это дает

$$\tau = \frac{(-1)^{n+1}}{n\pi}. \quad (7-14.66)$$

Погрешность приближения $y_n(\theta)$, таким образом, равна

$$\eta_n(\theta) = \sqrt{x} - y_n(\theta) = \frac{1}{n\pi} \left(\cos n\theta + n\theta \left[\text{si}(n\theta) - \frac{\pi}{2} \right] \right). \quad (7-14.67)$$

Чтобы получить точки максимума колеблющейся погрешности $\eta_n(\theta)$, приравняем производную нулю. Это приводит к условию

$$\text{si}(n\theta) = \frac{\pi}{2}.$$

Но в этих точках

$$\eta_n(\theta) = -\frac{1}{n\pi} \cos n\theta.$$

Следовательно,

$$\eta_n = \frac{1}{n\pi}$$

есть граница абсолютной величины погрешности во всем интервале. С возрастанием n величина η_n медленно убывает до нуля. Но даже $n=6$ дает точность до $\pm 0,053$; т. е. некоторый полином шестой степени аппроксимирует $\sqrt{x}$ с максимальной ошибкой $\pm 0,053$ во всем интервале между 0 и 1. Этот полином получается подстановкой значения τ из (66) в формулу (56)

$$\begin{aligned} y_6(x) &= \frac{1}{6\pi} \left[1 + 72x - \frac{840}{3}x^2 + \frac{3584}{5}x^3 - \frac{6912}{7}x^4 + \frac{6144}{9}x^5 - \frac{2048}{11}x^6 \right] = \\ &= 0,053 + 3,820x - 14,854x^2 + 38,028x^3 - 52,385x^4 + \\ &\quad + 36,217x^5 - 9,877x^6. \end{aligned}$$

Прежде чем мы занялись исследованием погрешности, у нас был уже метод получения τ ; мы воспользовались для этого условием (57). Так как ошибка в точке $x=1$ мала, мы можем ожидать, что оба найденных значения для τ будут лишь слегка отличаться между собой. Это дает следующее приближенное соотношение, которое

действительно даже при малых n с замечательной точностью и становится точным при $n = \infty$:

$$\sum_{m=0}^n \frac{c_n^m}{2m-1} = (-1)^{n+1} \pi n. \quad (7-14.68)$$

Например, при $n=6$ для π получается значение 3,1427, что означает точность до 0,001. Мы вернемся к этой задаче еще раз в следующем параграфе.

Пример 5. Следующий пример выбран для иллюстрации того, как τ -метод может служить для хорошего аппроксимирования функций в комплексной области. Функция

$$y = \operatorname{arctg} z \quad (7-14.69)$$

может быть определена дифференциальным уравнением

$$y' = \frac{1}{1+z^2}.$$

Положим

$$y = zu(z). \quad (7-14.70)$$

Тогда $u(z)$ есть *четная* функция от z и, таким образом, может быть представлена как функция от z^2 . Поэтому мы вводим новую переменную

$$z^2 = \xi \quad (7-14.71)$$

и получаем для $u(\xi)$ дифференциальное уравнение

$$2\xi u' + u = \frac{1}{1+\xi}. \quad (7-14.72)$$

Таким образом, нам надо решить уравнение

$$(1+\xi)(2\xi u' + u) = 1 + \tau T_n^*(\xi). \quad (7-14.73)$$

Бесконечный ряд Тейлора, получающийся при подстановке вместо $u(\xi)$ бесконечного полинома (без τ -члена) будет

$$S(\xi) = 1 - \frac{\xi}{3} + \frac{\xi^2}{5} - \dots + (-1)^k \frac{\xi^k}{2k+1} + \dots \quad (7-14.74)$$

Теперь мы образуем полиномы $Q_m(\xi)$, опуская первое уравнение, чтобы избавиться от переопределения. Поэтому построение $Q_m(\xi)$ начинаем со значения $m=1$ и получаем

$$Q_m(\xi) = (-1)^{m-1} S_{m-1}(\xi). \quad (7-14.75)$$

Следовательно, решением уравнения (73), если отвлечься от уравнения для коэффициентов при $m=0$, будет

$$u_{n-1}(\xi) = \tau \sum_{m=1}^n (-1)^{m-1} c_n^m S_{m-1}(\xi). \quad (7-14.76)$$

Опущенное уравнение

$$u(0) = 1 + \tau c_n^0 \quad (7-14.77)$$

дает для τ равенство

$$\tau = \frac{1}{\sum_{m=0}^n (-1)^{m-1} c_n^m} = \frac{1}{T_n^*(-1)}, \quad (7-14.78)$$

откуда

$$u_{n-1}(\xi) = \frac{\sum_{m=1}^n (-1)^m c_n^m S_{m-1}(\xi)}{T_n^*(-1)}. \quad (7-14.79)$$

Оценка максимальной погрешности дает

$$\eta_{n-1} = |\tau| = \frac{1}{|T_n^*(-1)|}. \quad (7-14.80)$$

Если подставить вместо ξ значение 1, ряд Тейлора дает знаменитый ряд Лейбница для $\frac{\pi}{4}$:

$$1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \dots,$$

который очень медленно сходится. Взяв частные суммы с весами чебышевских коэффициентов, согласно формуле (79), мы получаем быстро сходящийся ряд. Первые пять приближений для $\frac{\pi}{4}$ получаются:

по методу Тейлора

$$1; \frac{2}{3} = 0,667; \frac{13}{15} = 0,867; \frac{76}{105} = 0,724; \frac{789}{945} = 0,835;$$

по τ -методу

$$\begin{aligned} \frac{2 \cdot 1}{1+2} &= \frac{2}{3} = 0,667, \\ \frac{8 + 8 \frac{2}{3}}{1+8+8} &= \frac{40}{51} = 0,7843, \\ \frac{18 + 48 \frac{2}{3} + 32 \frac{13}{15}}{1+18+48+32} &= \frac{1166}{1485} = 0,78518, \\ \frac{32 + 160 \cdot \frac{2}{3} + 256 \frac{13}{15} + 128 \frac{76}{105}}{1+32+160+256+128} &= \frac{47584}{60585} = 0,785409, \\ \frac{50 + 400 \frac{2}{3} + 1120 \frac{13}{15} + 1280 \frac{76}{105} + 512 \frac{789}{945}}{1+50+400+1120+1280+512} &= \frac{2496018}{3178035} = \\ &= 0,7853966 \left(\frac{\pi}{4} = 0,785398163 \right). \end{aligned} \quad (7-14.81)$$

Однако мы имели в виду распространить наше решение на произвольные комплексные значения ξ . С этой целью мы воспользуемся «параметрическим методом», разобранным в конце § 13. Согласно формуле (13.20) единственное отличие этого метода состоит в том, что коэффициенты c_n^m должны быть заменены выражениями $c_n^m \xi^{-m}$. Следовательно, решение (76) теперь преобразуется в

$$u(\xi) = \tau \sum_{m=1}^n (-1)^{m-1} c_n^m \xi^{-m} S_{m-1}(\xi). \quad (7-14.82)$$

Опущенное уравнение, которое остается тем же: $u(0) = 1 + \tau c_n^0$, определяет τ в виде

$$\tau = \frac{1}{\sum_{m=0}^n (-1)^{m-1} c_n^m \xi^{-m}} = \frac{-1}{T_n^*\left(-\frac{1}{\xi}\right)}, \quad (7-14.83)$$

а окончательное решение получается в виде

$$u_{n-1}(\xi) = \frac{1}{T_n^*\left(-\frac{1}{\xi}\right)} \sum_{m=1}^n (-1)^m \frac{c_n^m}{\xi^m} S_{m-1}(\xi). \quad (7-14.84)$$

Переходя обратно к первоначальной переменной z , мы получаем, наконец, следующую последовательность приближений для функции $y = \operatorname{arctg} z$:

$$y_{n-1}(z) = \frac{z}{T_n^*\left(-\frac{1}{z^2}\right)} \sum_{m=1}^n (-1)^m \frac{c_n^m}{z^{2m}} S_{m-1}(z^2). \quad (7-14.85)$$

Это уже не простое полиномиальное приближение, а *отношение двух полиномов*. Например, полагая $n=4$, получаем следующее приближение:

$$\operatorname{arctg} z = \frac{32z}{105} \frac{420 + 700z^2 + 329z^4 + 38z^6}{128 + 256z^2 + 160z^4 + 36z^6 + z^8}, \quad (7-14.86)$$

Вычислим, например, при помощи этого приближения значения $\operatorname{arctg} z$ при $z=2$ и $z=\frac{i}{2}$. В первом случае получаем

$$\operatorname{arctg} 2 = \frac{64}{105} \frac{10916}{6016} = 1,10598 \quad (7-14.87)$$

при точном значении этой функции 1,10715.

Во втором случае получаем

$$\operatorname{arctg} \frac{i}{2} = \frac{16i}{105} \frac{67832}{18817} = 0,54930673i. \quad (7-14.88)$$

Но функция $\operatorname{arctg} z$ для мнимых значений аргумента определяется формулой

$$\operatorname{arctg} ip = \frac{i}{2} \log \frac{1+p}{1-p}, \quad (7-14.89)$$

следовательно,

$$\operatorname{arctg} \frac{i}{2} = \frac{i}{2} \lg 3 = 0,54930614i. \quad (7-14.90)$$

Какова сходимость приближения (85) при бесконечном возрастании n ? Погрешность пропорциональна τ , и, когда τ стремится к нулю, приближение неограниченно стремится к $f(z)$. Но в нашем примере

$$\tau = -\frac{1}{T_n^* \left(-\frac{1}{z^2} \right)}. \quad (7-14.91)$$

Полиномы Чебышева $T_n^*(x)$ обладают тем свойством, что они бесконечно возрастают, когда n стремится к бесконечности, в любой точке x , расположенной вне интервала $[0, 1]$. Внутри интервала они остаются ограниченными пределами ± 1 . Следовательно, сходимость будет сохраняться во всякой точке z , кроме тех, которые удовлетворяют условию

$$0 \leq -\frac{1}{z^2} \leq 1. \quad (7-14.92)$$

Следовательно, сходимость имеет место повсюду в комплексной плоскости, кроме точек мнимой оси. Но даже на мнимой оси исключаются лишь точки, лежащие вне интервала $(-i, i)$. Во всех точках между $-i$ и $+i$ сходимость все еще сохраняется.

Это как раз такой характер сходимости, который можно предсказать из теоретических соображений. Луч в комплексной плоскости, соединяющий точки 0 и z , не должен содержать особых точек. Следовательно, особая точка отбрасывает за собой тень, простирающуюся до бесконечности. Эта тень прочерчивается продолжением прямой, соединяющей начало с особой точкой. Особые точки функции $\operatorname{arctg} z$ суть $z = \pm i$. Следовательно, решение должно становиться расходящимся во всех точках мнимой оси, лежащих за этими точками.

15. Оценка погрешности τ -метода

В предыдущем параграфе сильная сходимость приближений, полученных τ -методом, была показана на численных примерах. Однако должна существовать теоретическая оценка погрешности, внутренне связанная с этими приближениями. Связь полиномов Чебышева с тригонометрическими функциями дает нам возможность развить простой алгебраический метод оценки погрешности решения линейного дифференциального уравнения, получаемого путем применения τ -метода.

Рассмотрим дифференциальное уравнение вида

$$A(x)y' + B(x)y + C(x) = \tau T_n^*(x). \quad (7-15.1)$$

Хотя эта форма уравнения имеет как будто весьма частный характер, она в действительности охватывает широкий класс задач. Кроме того, в случае дифференциального уравнения второго и высшего порядков целесообразно ввести дополнительные функции и преобразовать заданное дифференциальное уравнение в систему уравнений вида (1).

Таким образом, метод, который мы сейчас рассмотрим, применим в действительности к более широкой группе задач, которые могут быть приведены к двум или более уравнениям типа (1) и решаться по τ -методу. Для наших целей достаточно будет рассмотреть лишь одно уравнение вида (1).

Правая часть присоединена к уравнению (1) для того, чтобы иметь возможность решить уравнение с помощью конечного разложения. Вообще говоря, нам может понадобиться более одного τ -члена в правой части; однако мы можем оценить ошибку, соответствующую каждому τ -члену отдельно, и затем взять сумму их абсолютных величин. Ввиду этого достаточно будет провести наше исследование только для одного τ -члена.

Точное решение заданного дифференциального уравнения не требует в правой части никакого поправочного члена:

$$A(x)y' + B(x)y + C(x) = 0. \quad (7-15.2)$$

Разность между точным решением $y(x)$ и приближенным $y_n(x)$ представляет погрешность решения

$$\eta_n(x) = y(x) - y_n(x), \quad (7-15.3)$$

удовлетворяющую дифференциальному уравнению

$$A(x)\eta_n'(x) + B(x)\eta_n(x) = -\tau T_n^*(x). \quad (7-15.4)$$

Нашей целью будет найти приближенное решение этого дифференциального уравнения и, таким образом, получить оценку погрешности $\eta_n(x)$.

Введем в качестве новой переменной угловую переменную θ , заменив x через

$$x = \cos^2 \frac{\theta}{2}; \quad dx = -\frac{1}{2} \sin \theta d\theta. \quad (7-15.5)$$

Тогда мы получаем относительно новой переменной следующее дифференциальное уравнение:

$$A_1\eta' + B_1\eta = -\tau \cos n\theta,$$

где

$$A_1(\theta) = -\frac{2A \left(\cos^2 \frac{\theta}{2} \right)}{\sin \theta}, \quad (7-15.6)$$

$$B_1(\theta) = B \left(\cos^2 \frac{\theta}{2} \right). \quad (7-15.7)$$

Правую часть можно заменить через $-\tau e^{in\theta}$, условившись при этом, что мы используем лишь *вещественную* часть решения.

Но уравнение

$$A_1 \eta' + B_1 \eta = -\tau e^{in\theta} \quad (7-15.8)$$

в предположении, что A_1 и B_1 суть постоянные, можно рассматривать как дифференциальное уравнение баллистического гальванометра, причем правой части соответствует внешняя периодическая сила, которая сообщает гальванометру вынужденные колебания. Решением уравнения (8) является

$$\eta(\theta) = \frac{-\tau}{inA_1 + B_1} e^{in\theta}. \quad (7-15.9)$$

Это означает, что гальванометр следует внешней силе с измененными амплитудой и фазой. В нашем случае A_1 и B_1 являются не постоянными, а заданными функциями от θ . Тем не менее по сравнению с быстро изменяющейся показательной функцией $e^{in\theta}$ они ведут себя почти как постоянные. Поэтому мы не слишком ошибемся и получим решение для $\eta_n(\theta)$ по крайней мере для достаточно больших значений n , если для нашей цели воспользуемся решением (9), хотя мы понимаем, что это решение является только приближением. Для точного решения мы должны были бы положить

$$\eta = -\mathcal{G}(\theta) \tau e^{in\theta} \quad (7-15.10)$$

и получить для переменной амплитуды $\mathcal{G}(\theta)$ дифференциальное уравнение

$$A_1(\mathcal{G}' + in\mathcal{G}) + B_1\mathcal{G} = 1. \quad (7-15.11)$$

Однако для достаточно большого n мы можем считать $\mathcal{G}'$ пренебрежимо малым сравнительно с $in\mathcal{G}$. Тогда мы снова приходим к решению (9).

Как видим, погрешность $\eta_n(\theta)$ нашего решения будет иметь *периодический* характер. Если записать (9) в полярной форме

$$\eta_n(\theta) = \frac{-\tau}{\sqrt{B_1^2 + n^2 A_1^2}} e^{i(n\theta - \varphi)}, \quad (7-15.12)$$

где

$$\operatorname{tg} \varphi = \frac{B_1}{nA_1}, \quad (7-15.13)$$

то, взяв вещественную часть выражения (12), получим

$$\eta_n(\theta) = -\frac{\tau}{\sqrt{B_1^2 + n^2 A_1^2}} \cos(n\theta - \varphi). \quad (7-15.14)$$

Это есть гармоническое колебание с переменной фазой и амплитудой.

Из выражения (14) можно вывести некоторые заключения. В первую очередь нам надо знать, какова будет *максимальная погрешность*, с которой можно встретиться в *любой* точке интервала. С этой целью мы исследуем величину

$$B^2(x) + \frac{n^2 A^2(x)}{x(1-x)} = \varphi^2(x) \quad (7-15.15)$$

и найдем ее минимум в интервале $[0,1]$. Пусть этот минимум будет $\varphi_{\min}$. Тогда максимальная возможная погрешность $\eta_{\max}$ определяется соотношением

$$\eta_n^0 \leq |\eta_{\max}| \leq \frac{\tau}{\varphi_{\min}}. \quad (7-15.16)$$

Далее, «метод вынужденных колебаний», выраженный формулой (9), даст достаточно близкую оценку погрешности $\eta(\theta)$ в любой точке $x = \cos^2\left(\frac{\theta}{2}\right)$. Однако мы наталкиваемся на затруднение в двух конечных точках интервала $x=0$ и $x=1$. Здесь знаменатель выражения $A_1(\theta)$ [ср. (6)] делается равным нулю и, следовательно, $A_1(\theta)$ становится бесконечно большим. Оценка погрешности (9) окажется поэтому равной нулю. Это указывает, что ошибка в точках $x=0$ и $x=1$ особенно мала. Однако именно в конечной точке $x=1$ погрешность представляет особенный интерес, так как при параметрическом методе переменная точка z оказывается конечной точкой интервала.

Во многих задачах трудность возникает только в точке $x=1$, ибо в точке $x=0$ функция может обращаться в нуль ввиду наличия множителя x . Тогда оценка погрешности (8) остается действительной в нижней граничной точке и требует уточнения лишь в верхней граничной точке.

Для получения оценки при $x=1$ мы освободимся от знаменателя в дроби (6) путем умножения на $\sin \theta$. Тогда правая часть уравнения принимает вид

$$-\tau \cos n\theta \sin \theta = -\frac{\tau}{2} [\sin(n+1)\theta - \sin(n-1)\theta],$$

и тот же метод, который привел к решению (9), теперь дает

$$\eta_n = \frac{-\tau}{-2i(n+1)A + B \sin \theta} \frac{e^{i(n+1)\theta}}{2i} + \frac{\tau}{-2i(n-1)A + B \sin \theta} \frac{e^{i(n-1)\theta}}{2i}. \quad (7-15.17)$$

В точке $x=1$, т. е. при $\theta=0$, получаем

$$\eta_n(1) = \frac{\tau}{2} \left(-\frac{1}{2(n+1)A(1)} + \frac{1}{2(n-1)A(1)} \right) = \frac{\tau}{2(n^2-1)A(1)}. \quad (7-15.18)$$

Эта формула показывает, что, тогда как, вообще, погрешность есть величина порядка $\frac{\tau}{n}$, в конечной точке интервала порядок величины снижается до $\frac{\tau}{n^2}$.

В свете этих результатов мы еще раз рассмотрим примеры предыдущего параграфа.

Пример 1. Здесь

$$A(x) = -1, \quad B(x) = 1.$$

Верхняя граница (15.16) для погрешности в любой точке интервала равна

$$\eta_n^0 = \frac{|\tau_n|}{\sqrt{4n^2+1}},$$

тогда как оценка ошибки в конечной точке $x=1$ будет

$$\eta_n(1) = -\frac{\tau_n}{2(n^2-1)}. \quad (7-15.19)$$

Но τ_n в этом примере все положительны, и следовательно, кажется непонятным тот факт, что значения для e , данные в (14.8), приближаются к пределу попеременно сверху и снизу. Формула (19) указывает, что погрешность должна бы оставаться постоянно отрицательной. Кажущаяся невязка объясняется следующим образом. Оценка (17) дает определенную погрешность в точке $x=0$, т. е. $\theta=\pi$. Эта погрешность равна погрешности при $x=1$, умноженной на $(-1)^{n+1}$:

$$\eta_n(0) = (-1)^{n+1} \eta_n(1). \quad (7-15.20)$$

Но в решении (14.6) мы в точности удовлетворили краевому условию (14.3), а это означает, что мы не имеем ошибки при $x=0$. Ввиду этого мы должны скорректировать нашу оценку периодической погрешности «систематической ошибкой», обусловленной тем фактом, что в точке $x=0$ мы не приняли во внимание естественную ошибку, которая должна бы существовать в этой точке.

Решая дифференциальное уравнение (4), мы не учли решения однородного дифференциального уравнения, которое, с постоянным множителем C , может быть всегда добавлено. Так как в нашей задаче однородное дифференциальное уравнение имеет решение e^x , то добавление этого решения изменит $\eta(1)$ на дополнительный член Ce , в то время как $\eta_n(0)$ изменится на C . Условие $\eta_n(0)=0$ определяет C :

$$C = (-1)^n \eta_n(1),$$

и таким образом, окончательная погрешность в точке $x=1$

$$\eta_n(1) = -\frac{\tau_n}{2(n^2-1)} [1 + (-1)^n e]. \quad (7-15.21)$$

Так как e больше единицы, то погрешность $\eta(1)$ будет знакопеременной: отрицательной при четном n и положительной при нечетном n , как это имеет место фактически. Оценка (21) весьма удовлетворительна, начиная со значения $n=3$.

Пример 2. Этот пример чисто алгебраический. Дифференциальный оператор отсутствует, так что уравнение (4) оказывается разрешимым точно. Вместо *оценки* ошибки $\eta_n(x)$ мы имеем точное алгебраическое *представление* ее. Сравнение с погрешностью разложения Тейлора было уже раньше рассмотрено и не нуждается в дальнейшем разборе.

Пример 3. Здесь

$$A(x) = x^2, \quad B(x) = 1 + x.$$

Погрешность оказывается наибольшей при $x=0$; оценка максимальной погрешности, таким образом, имеет вид

$$\eta_n^0 = |\tau_n|,$$

тогда как в конечной точке $x=1$ оценка получается следующей:

$$\eta_n(1) = \frac{\tau_n}{2(n^2-1)}. \quad (7-15.22)$$

Выражение (14.47) для τ_n показывает, что τ_n отрицательно для четного n и положительно для нечетного. Применяя формулу (22) к последовательным приближениям (14.50) значения $y(1)$ и учитывая, что τ_n соответствует приближению $y_{n-1}(x)$, находим, что оценка (22) становится эффективной, начиная с $n=5$. Начиная с этого значения, попеременное приближение снизу и сверху не нарушается.

Пример 4. В этом примере

$$A(x) = 2x, \quad B(x) = -1,$$

и для оценки верхней границы погрешности в любой точке интервала получаем

$$\eta_n^0 = |\tau_n|,$$

а оценка погрешности в конечной точке дает

$$\eta_n(1) = \frac{\tau_n}{4(n^2-1)}. \quad (7-15.23)$$

Погрешность в точке $x=1$ столь мала в сравнении с гораздо большей погрешностью в точке $x=0$, что мы очень мало теряем в точности, если положим $\eta(1)=0$ и определим τ из граничного условия

$y(1) = 1$, как мы и поступили в равенстве (14.57). Однако интересно продемонстрировать точность оценки (23), приняв ее в расчет и определив τ из условия

$$\tau_n \sum_{m=0}^n \frac{c_n^m}{2m-1} = 1 - \frac{\tau_n}{4(n^2-1)}. \quad (7-15.24)$$

Тогда мы получаем для величины $\frac{1}{n\tau_n}$ значение

$$\frac{1}{n\tau_n} = \frac{1}{4n(n^2-1)} + \frac{1}{n} \sum_{m=0}^n \frac{c_n^m}{2m-1}. \quad (7-15.25)$$

Согласно рассуждениям относительно погрешности в § 14 эта величина теоретически должна бы равняться $(-1)^{n+1}\pi$. Ранее мы получили для случая $n=6$ [ср. (14.68)] значение $\pi=3,1427$, не учитывая первый член в правой части формулы (25). Но если принять в расчет этот член, то получаем гораздо более точное значение $\pi=3,141522$. Точность возрастает от 1:2800 до 1:45 000.

Пример 5. В этом примере был применен параметрический метод, и функция $y=f(z^2)$ была получена таким образом, что z^2 рассматривалась как простая постоянная данного дифференциального уравнения. Здесь

$$A(x) = (1 + z^2 x) 2x, \quad B(x) = 1 + z^2 x.$$

Общая точность во всем интервале $[0,1]$ здесь не играет роли, так как мы используем решение только в точке $x=1$, которая в первоначальных переменных соответствует точке $x=z$. Мы, следовательно, получаем

$$\eta_n(z) = \frac{\tau(z)}{4(n^2-1)(1+z^2)}.$$

Эта формула действительна для функции $u(z)$. Для функции $\operatorname{arctg} z$ мы получаем, согласно равенству (14.70),

$$\eta_n(z) = \frac{z \tau_n(z)}{4(n^2-1)(1+z^2)} = \frac{-z}{4(n^2-1)(1+z^2)} \cdot \frac{1}{T_n^* \left(-\frac{1}{z^2} \right)}. \quad (7-15.26)$$

Применим эту оценку погрешности к случаю $z=1$, т. е. к значениям $\frac{\pi}{4}$ в таблице (14.81). Величина $T_n(-1)$ меняет попеременно знак с положительного при четном n на отрицательный при нечетном. Поэтому последовательные приближения для $\frac{\pi}{4}$ должны подходить к точному значению снизу для нечетного n и сверху для четного. Это оправдывается, начиная со значения $n=3$. Напротив, погрешность

таблицы (14.87) плохо следует оценке по формуле (26). Если положить $z=2$, то $A(x)$ — уже недостаточно гладкая функция для того, чтобы формулу (26) можно было применить при $n=4$. Для этой цели нам нужно было бы перейти к несколько большим значениям n . Однако совершенно другое положение получается при подстановке $z = \frac{i}{2}$. Поразительно большая точность приближения (14.88) объясняется большим значением $T_4^*(4)$:

$$T_4^*(4) = 18\,871.$$

Здесь оценка (26) дает

$$\eta_4(z) = -\frac{i}{90 \cdot 18\,871} = -0,59 \cdot 10^{-6}i$$

в хорошем согласии с действительной ошибкой.

16. Квадратный корень из комплексного числа

Квадратный корень из комплексного числа можно получить обычным алгебраическим путем, но численная процедура, с которой это связано, весьма трудоемка. Нижеследующий метод дает результат гораздо быстрее.

Обозначим комплексное число, квадратный корень которого отыскивается, через $A + Bi$. Если A больше по модулю, чем B , то полагаем

$$\sqrt{A + Bi} = \sqrt{A} \sqrt{1 + \frac{Bi}{A}}. \quad (7-16.1)$$

Если, наоборот, B больше по модулю, чем A , то полагаем

$$\sqrt{A + Bi} = \sqrt{B} \sqrt{\frac{A}{B} + i}. \quad (7-16.2)$$

Следовательно, задача сводится к вычислению одной из следующих функций:

$$y = \sqrt{1 + ix} \quad (7-16.3)$$

или

$$y = \sqrt{i + x}, \quad (7-16.4)$$

где x ограничен интервалом $[0, 1]$ (функция (4) сводится к функции (3), если вынести множитель i за знак радикала). Условимся, что x может изменяться лишь между 0 и 1. Следовательно, $\sqrt{1 - ix}$ не может быть получено из выражения (3) подстановкой отрицательного значения x , а лишь заменой i на $-i$.

Нашей целью будет получить быстро сходящееся разложение для функции (3) с помощью τ -метода. С этой целью мы определяем

функцию (3) дифференциальным уравнением

$$\frac{y'}{y} = \frac{i}{2(1+ix)}. \quad (7-16.5)$$

Мы можем, однако, считать также, что решается дифференциальное уравнение

$$\frac{y'}{y} = \frac{1}{2(1+x)} \quad (7-16.6)$$

или

$$2(1+x)y' - y = 0$$

вдоль мнимой оси.

Согласно обычной процедуре, мы вводим в правую часть дифференциального уравнения (6) τ -член. Полином $T_n^*(x)$ оказывается автоматически пригодным для мнимой оси, если записать его в виде $T_n^*\left(\frac{x}{i}\right)$. Но в нашей задаче член y' является доминирующим. Поэтому мы получим лучшие результаты, если воспользуемся не прямо полиномами T_n^* , а их производными, формулируя нашу задачу следующим образом:

$$2(1+x)y' - y = \tau T_{n+1}^{*'}\left(\frac{x}{i}\right). \quad (7-16.7)$$

Разложение Тейлора в нашем примере дает обыкновенный биномиальный ряд

$$S(x) = 1 + \frac{x}{2} - \frac{x^2}{8} + \frac{x^3}{16} - \dots, \quad (7-16.8)$$

а метод канонических полиномов Q показывает, что частные суммы этого разложения должны быть взяты со следующими весами:

$$y_n(x) = \tau \sum_{m=0}^n \frac{c_n^m}{a_m(2m-1)} S_m(x). \quad (7-16.9)$$

Здесь c_n^m суть коэффициенты полинома, стоящего в правой части уравнения (7). Кроме того, a_m есть последний коэффициент частной суммы $S_m(x)$. Наконец, так как переменная x предполагается движущейся вдоль мнимой оси, x должен быть заменен через $\frac{x}{i}$.

В качестве примера примем $n=2$. Тогда

$$\begin{aligned} T_3^*(x) &= -1 + 18x - 48x^2 + 32x^3, \\ T_3^{*'}(x) &= 18 - 96x + 96x^2 \end{aligned}$$

и

$$T_3^{*'}\left(\frac{x}{i}\right) = 18 + 96ix - 96x^2.$$

Далее,

$$a_0 = 1, \quad a_1 = \frac{1}{2}, \quad a_2 = -\frac{1}{8}.$$

Следовательно,

$$\begin{aligned} y_2(x) &= \tau \left[-18 + 2 \cdot 96i \left(1 + \frac{ix}{2} \right) + \frac{8 \cdot 96}{2} \left(1 + \frac{ix}{2} + \frac{x^2}{8} \right) \right] = \\ &= \tau [238 + 192i + (-96 + 128i)x + 32x^2]. \end{aligned}$$

Множитель τ может быть определен из граничного условия

$$y(0) = 1,$$

которое дает

$$\tau = \frac{1}{238 + 192i}.$$

Окончательный результат будет

$$\begin{aligned} \sqrt{1 + ix} &= 1 + 0,0184x + 0,0810x^2 + i(0,5201x - 0,0653x^2) = \\ &= 1 + x(0,0184 + 0,0810x) (\pm 0,002) + \\ &\quad + ix(0,5201 - 0,0653x) (\pm 0,002i). \quad (7-16.10) \end{aligned}$$

Точность этого квадратичного приближения замечательно высока, ибо максимальная ошибка в любой точке интервала $[0,1]$ не превышает двух единиц в третьем десятичном разряде. Такая точность будет достаточной во многих задачах.

Если сделаем шаг дальше и выведем соответствующие формулы для случая $n=3$, то получим кубическое приближение, которое будет уже точно до двух единиц *четвертого* десятичного разряда (как в вещественной, так и мнимой части):

$$\begin{aligned} \sqrt{1 + ix} &= 1 - 0,00316x + 0,14237x^2 - 0,04079x^3 + \\ &\quad + i(0,50637x - 0,03108x^2 - 0,02020x^3). \quad (7-16.11) \end{aligned}$$

Квадратичное приближение для функции $\sqrt{i+x}$ будет

$$\begin{aligned} \sqrt{i+x} &= 0,7071 + x(0,3807 + 0,0111) + \\ &\quad + i[0,7071 - x(0,3548 - 0,1035x)]. \quad (7-16.12) \end{aligned}$$

Для удобства добавим еще две другие формулы на случай любой комбинации знаков у A и B :

$$\begin{aligned} \sqrt{-1 + ix} &= x(0,5201 - 0,0653x) + \\ &\quad + i[1 + x(0,0184 + 0,0810x)], \quad (7-16.13) \end{aligned}$$

$$\begin{aligned} \sqrt{i-x} &= 0,7071 - x(0,3548 - 0,1053x) + \\ &\quad + i[0,7071 + x(0,3807 + 0,0111x)]. \quad (7-16.14) \end{aligned}$$

В виде примера вычислим выражение $\sqrt{-3-8i}$:

$$\sqrt{-3-8i} = \sqrt{8} \sqrt{-i-0,375}.$$

Здесь подходит формула (14) при $x=0,375$ с заменой i на $-i$:

$$\begin{aligned} \sqrt{-3-8i} &= 2,8284 [0,7071 - 0,375 \cdot 0,3150 - \\ &\quad - i(0,7071 + 0,375 \cdot 0,3849)] = 1,6659 - 2,4081i. \end{aligned}$$

Точным решением является

$$\sqrt{-3-8i} = 1,66493 - 2,40250i.$$

Ошибка, следовательно, не превышает $0,3\%$.

17. Обобщение τ -метода. Метод избранных точек

τ -метод непосредственно применим лишь в случае, когда коэффициенты заданного линейного дифференциального уравнения суть полиномы относительно x . Для более общих уравнений мы не можем построить надлежащий поправочный член в правой части уравнения, который приводил бы к решению в виде конечного степенного ряда с быстрой сходимостью. Однако мы можем иначе сформулировать смысл τ -метода таким образом, чтобы сделать его пригодным к гораздо более широкому классу задач. Вернемся снова к основной идее τ -метода. Мы сделали линейное дифференциальное уравнение

$$Dy = 0 \quad (7-17.1)$$

разрешимым с помощью конечного степенного ряда путем введения в правую часть надлежаще выбранного поправочного члена. Иногда оказывается достаточным член вида $\tau T_n^*(x)$, но, вообще говоря, нам надо было положить

$$Dy = T_n^*(x) (\tau_1 + \tau_2 x + \dots + \tau_m x^{m-1}). \quad (7-17.2)$$

При помощи этого приема мы заменяем $y(x)$ некоторым рядом

$$y_p(x) = b_0 + b_1 x + \dots + b_p x^p, \quad (7-17.3)$$

который хорошо аппроксимирует $y(x)$ в интервале $[0,1]$. Применение этого процесса, однако, ограничивается линейными дифференциальными уравнениями с многочленными коэффициентами.

Теперь мы подойдем к этой задаче с совершенно другой точки зрения. Мы снова примем, что $y(x)$ должна быть аппроксимирована конечным многочленом степени p . Это дает нам возможность распорядиться $p+1$ постоянными

$$b_0, b_1, \dots, b_p, \quad (7-17.4)$$

которые однозначно определяются $p+1$ независимыми уравнениями. Это сразу указывает на то, что мы не можем рассчитывать на то, чтобы удовлетворить данному дифференциальному уравнению в каком-нибудь непрерывном интервале переменной x . Если мы удовлетворим этому уравнению и заданным граничным условиям только в $p+1$ точках, то этого, вообще говоря, достаточно для однозначного определения постоянных b и всякие дополнительные условия привели бы к переопределению.

Вопрос теперь сводится к надлежащему выбору $p+1$ точек, в которых должно удовлетворяться заданное уравнение. Если даны начальные значения или граничные условия, то некоторые из коэффициентов b_i уже определяются этими условиями и число свободных параметров равно уже не $p+1$, а $p+1-\nu$, где ν — число заданных граничных условий. Наша задача, таким образом, сводится к надлежащему выбору $n=p+1-\nu$ точек. Одним из возможных выборов было бы сосредоточение всех этих n точек в ближайшей окрестности начала, т. е. удовлетворение заданным уравнениям при значениях x :

$$x=0, \epsilon, 2\epsilon, \dots, (n-1)\epsilon,$$

где ϵ стремится к нулю. Полученное таким образом разложение соответствует разложению Тейлора для функции $y=f(x)$, которое обрывается на $(p+1)$ -м члене.

Совсем другое распределение точек подсказывается равенством (2). Правая часть дифференциального уравнения (2) обращается в нуль в нулевых точках полинома $T_n^*(x)$. Эти точки суть

$$x_k = \frac{1 + \cos \frac{(2k-1)\pi}{2n}}{2} \quad (k=1, 2, \dots, n). \quad (7-17.5)$$

Если оказывается целесообразным взять $T_{n+1}^{*'}(x)$ вместо $T_n^*(x)$, то вместо точек (5) нужны будут нули «полиномов Чебышева второго рода», которые определяются из формул

$$x_k = \frac{1 + \cos \frac{k\pi}{n+1}}{2} \quad (k=1, 2, \dots, n). \quad (7-17.6)$$

Эти точки имеют простое геометрическое истолкование. На отрезке $[0,1]$, как на диаметре, строим полуокружность с центром в точке $x=\frac{1}{2}$ и радиусом, равным $\frac{1}{2}$; делим эту полуокружность на $n+1$ равных частей и точки деления проектируем на диаметр. Исключив обе концевые точки¹⁾, мы получаем, таким образом, n неравномерно располо-

¹⁾ Мы не очень много потеряем в точности, если включим и две концевые точки, заменив в формуле (6) $n+1$ через $n-1$. Вычислительная схема часто упрощается при включении этих двух концевых точек.

женных точек, рассеянных по всему интервалу, а не собранных вблизи точки $x=0$. При таком распределении нулей мы получим полиномиальное приближение, погрешность которого колеблется практически одинаково по всему заданному интервалу, вместо того, чтобы давать очень малую погрешность в окрестности точки $x=0$, но большую погрешность в окрестности точки $x=1$.

Теперь наше неоднородное дифференциальное уравнение (2) можно решить двумя различными путями. Первый путь — это решить рекуррентные соотношения для коэффициентов и таким образом определить коэффициенты b_i и множители τ_i . Другой путь — это вовсе опустить множители τ_i , выполнить дифференциальную операцию D над конечным рядом (3) и приравнять нулю полученное выражение во всех тех точках, в которых правая часть дифференциального уравнения обращается в нуль. Оба метода дают тождественные результаты, но второй метод имеет более универсальное значение. Он применим к линейным дифференциальным уравнениям, коэффициенты которых не обязательно алгебраические выражения относительно x . Он в равной степени применим к интегральным уравнениям или смешанным дифференциально-интегральным уравнениям, если они линейны. Общая процедура всегда одна и та же. Мы заменяем $y(x)$ конечным многочленным разложением, выполняем над этим разложением операции, указанные заданным дифференциальным или интегральным оператором, и приравниваем результат операции нулю не для произвольных значений x , что невозможно, а для тщательно выбранного дискретного множества значений x , которое обеспечивает удовлетворительно одинаковое распределение погрешности по всему заданному интервалу. При этом мы получаем систему обыкновенных линейных алгебраических уравнений относительно коэффициентов b_k , из которых они могут быть однозначно определены. Если степень p аппроксимирующего многочлена изменяется, весь процесс должен быть заново повторен, причем получается совершенно новая система значений для коэффициентов.

Этот метод решения дифференциальных или интегральных уравнений с помощью хорошо сходящихся степенных рядов является обобщением метода «тригонометрической интерполяции», рассмотренного ранее в гл. IV, 16. Если известны значения функции $y=f(x)$, то близкое полиномиальное приближение для $f(x)$ можно получить, сделав значения полинома равными значениям функции в точках (5) или (6). Это сводится к равноотстоящей тригонометрической интерполяции функции от θ с помощью функций Фурье $\cos k\theta$. Если, однако, значения функции $f(x)$ неизвестны, но мы знаем закон, определяющий $f(x)$ с помощью линейного оператора, то мы также можем получить хорошо сходящееся полиномиальное приближение для $f(x)$ путем интерполирования значений оператора вместо значений функции. Процесс в обоих случаях сходен; единственное отличие состоит в том, что в последнем случае результирующая система линейных уравнений не обладает свойством ортогональности, которое давало бы

возможность их явного решения. Так как, однако, во многих практических приложениях степень аппроксимирующих полиномов может не превышать 4 или 5, то получающаяся линейная система будет небольшого порядка и может быть решена последовательным исключением. Мы, таким образом, получаем изящную численную схему, которая обобщает τ -метод на гораздо более широкий класс задач¹⁾.

ЛИТЕРАТУРА К ГЛАВЕ VII

- [1] Bromwich Th. J., *I'a*, Introduction to the Theory of Infinite Series (Macmillan, London, 1926).
- [1a] Бари Н. К., Тригонометрические ряды, Физматгиз, М., 1961.
- [2] Е. Янке и Ф. Эмде, Таблицы функций с формулами и кривыми, Физматгиз, М., 1959.
- [2a] Сегал Б. И. и Семендяев К. А., Пятизначные математические таблицы, Физматгиз, М., 1959.
- [3] Szegő G., Orthogonal Polynomials, Am. Math. Soc. Colloq. Pub. 23, 1939.
- [3a] Геронимус Я. Л., Теория ортогональных многочленов, Гостехиздат, М., 1950.
- [4] Tables of the Chebyshev Polynomials $S_n(x)$ and $C_n(x)$, AMS 9, National Bureau of Standards (1952).

Статьи:

- [5] Lanczos C., Trigonometric Interpolation of Empirical and Analytical Functions, J. Math. Phys. 17, 123 (1938).
- [6] Miller J. C. P., Two Numerical Applications of Chebyshev Polynomials, Roy. Soc. Edin. Proc. 62, 204 (1946).

¹⁾ Практический пример этого метода избранных точек см. в [5], стр. 96.

ЧИСЛОВЫЕ ТАБЛИЦЫ

Т а б л и ц а I
Амплитуды комплексных корней (см. стр. 53)

<i>r</i>	<i>A (r)</i>	<i>r</i>	<i>A (r)</i>	<i>r</i>	<i>A (r)</i>	<i>r</i>	<i>A (r)</i>
0.80	3.3588	0.90	4.7774	1.00	7.0669	1.10	10.8412
81	4732	91	9594	01	3644	11	11.3353
82	5928	92	5.1505	02	6773	12	8553
83	7179	93	3510	03	8.0062	13	12.4028
84	8490	94	5615	04	3522	14	9790
85	9862	95	7825	05	7161	15	13.5857
86	4.1299	96	6.0146	06	9.0990	16	14.2243
87	2805	97	2585	07	5019	17	8967
88	4383	98	5147	08	9257	18	15.6046
89	6038	99	7839	09	10.3718	19	16.3499
						20	17.1345

Т а б л и ц а II
Матрицы преобразования *B_n* (см. стр. 54)
(Подчеркнутые числа — отрицательные.)

<i>B₁</i>	<i>B₂</i>	<i>B₃</i>	<i>B₄</i>
$\begin{bmatrix} \underline{1} & 1 \\ 1 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & \underline{1} & 1 \\ \underline{2} & 0 & 2 \\ 1 & 1 & 1 \end{bmatrix}$	$\begin{bmatrix} \underline{1} & 1 & \underline{1} & 1 \\ 3 & \underline{1} & \underline{1} & 3 \\ 3 & \underline{1} & 1 & 3 \\ \underline{1} & 1 & 1 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & \underline{1} & 1 & \underline{1} & 1 \\ \underline{4} & 2 & 0 & \underline{2} & 4 \\ 6 & 0 & \underline{2} & 0 & 6 \\ \underline{4} & \underline{2} & 0 & 2 & 4 \\ \underline{1} & 1 & 1 & 1 & 1 \end{bmatrix}$

Продолжение табл. II

$$B_5$$

1	1	1	1	1	1
5	3	1	1	3	5
10	2	2	2	2	10
10	2	2	2	2	10
5	3	1	1	3	5
1	1	1	1	1	1

$$B_6$$

1	<u>1</u>	1	<u>1</u>	1	<u>1</u>	1
<u>6</u>	4	<u>2</u>	0	2	<u>4</u>	6
15	<u>5</u>	<u>1</u>	3	<u>1</u>	<u>5</u>	15
<u>20</u>	0	4	0	<u>4</u>	0	20
15	5	<u>1</u>	3	<u>1</u>	5	15
<u>6</u>	<u>4</u>	<u>2</u>	0	2	4	6
1	1	1	1	1	1	1

$$B_7$$

1	1	1	1	1	1	1	1
7	5	3	1	1	3	5	7
21	9	1	3	3	1	9	21
35	5	5	3	3	5	5	35
35	5	5	3	3	5	5	35
21	9	1	3	3	1	9	21
7	5	3	1	1	3	5	7
1	1	1	1	1	1	1	1

$$B_8$$

1	<u>1</u>	1	<u>1</u>	1	<u>1</u>	1	<u>1</u>	1
<u>8</u>	6	<u>4</u>	2	0	<u>2</u>	4	<u>6</u>	8
28	<u>14</u>	4	2	<u>4</u>	2	4	<u>14</u>	28
<u>56</u>	14	4	<u>6</u>	0	6	<u>4</u>	<u>14</u>	56
70	0	<u>10</u>	0	6	0	<u>10</u>	0	70
<u>56</u>	<u>14</u>	4	6	0	<u>6</u>	<u>4</u>	14	56
28	14	4	<u>2</u>	<u>4</u>	<u>2</u>	4	14	28
<u>8</u>	<u>6</u>	<u>4</u>	<u>2</u>	0	2	4	6	8
1	1	1	1	1	1	1	1	1

[illegible]

$$B_{10}$$
[illegible]
$$B_{11}$$
[illegible]

B_{12}

1	1	1	1	1	1	1	1	1	1	1	1	1
12	10	8	6	4	2	0	2	4	6	8	10	12
66	44	26	12	2	4	6	4	2	12	26	44	66
220	110	40	2	12	10	0	10	12	2	40	110	220
495	165	15	27	17	5	15	5	17	27	15	165	495
792	132	48	36	8	20	0	20	8	36	48	132	792
924	0	84	0	28	0	20	0	28	0	84	0	924
792	132	48	36	8	20	0	20	8	36	48	132	792
495	165	15	27	17	5	15	5	17	27	15	165	495
220	110	40	2	12	10	0	10	12	2	40	110	220
66	44	26	12	2	4	6	4	2	12	26	44	66
12	10	8	6	4	2	0	2	4	6	8	10	12
1	1	1	1	1	1	1	1	1	1	1	1	1

Т а б л и ц а III
Обращение собственных значений (см. стр. 213)

Наименьшее значение v : $v_0 = \frac{\sqrt{3}}{\pi} = 0,5513288953$. Для очень малых u :
 $u = u_0^* \sqrt{v - v_0}$, где $u_0^* = \sqrt{\frac{5\sqrt{3}}{\pi}} = 1,660314572$. Взятое из таблицы u^* , заключающееся между $v = 0,56$ и $v = 1$, следует умножить на $\sqrt{v - v_0}$. Поэтому $u^2 = u^{*2} (v - 0,55133)$. При $v = 1$ таблица дает непосредственно значение u .

v	u^*	v	u^*	v	u^*	v	u^*
0.56	1.6548	0.67	1.5945	0.78	1.5489	0.89	1.5163
57	6485	68	5898	79	5457	90	5138
58	6425	69	5853	80	5427	91	5113
59	6366	70	5810	81	5395	92	5090
60	6309	71	5767	82	5363	93	5067
61	6253	72	5724	83	5331	94	5045
62	6198	73	5684	84	5301	95	5024
63	6145	74	5644	85	5272	96	5004
64	6093	75	5606	86	5243	97	4984
65	6043	76	5568	87	5216	98	4965
66	5993	77	5527	88	5189	99	4947
						1.00	4929

v	u	v	u	v	u	v	u
1.00	1.0000	2.00	2.0000	3.00	3.0000	4.00	4.0000
02	0199	02	0200	02	0200	02	0200
04	0393	04	0397	04	0398	04	0398
06	0584	06	0592	06	0595	06	0596
08	0772	08	0785	08	0790	08	0793

Продолжение табл. III

<i>v</i>	<i>u</i>	<i>v</i>	<i>u</i>	<i>v</i>	<i>u</i>	<i>v</i>	<i>u</i>
1.10	1.0958	2.10	2.0978	3.10	3.0985	4.10	4.0989
12	1144	12	1169	12	1180	12	1184
14	1336	14	1359	14	1372	14	1379
16	1508	16	1548	16	1564	16	1573
18	1688	18	1737	18	1757	18	1767
20	1867	20	1925	20	1948	20	1960
22	2046	22	2113	22	2139	22	2154
24	2224	24	2300	24	2330	24	2346
26	2405	26	2488	26	2521	26	2539
28	2583	28	2676	28	2712	28	2732
30	2763	30	2863	30	2904	30	2926
32	2944	32	3052	32	3096	32	3119
34	3125	34	3241	34	3287	34	3313
36	3308	36	3431	36	3480	36	3507
38	3492	38	3622	38	3674	38	3702
40	3680	40	3813	40	3868	40	3898
42	3868	42	4007	42	4064	42	4095
44	4059	44	4201	44	4260	44	4292
46	4252	46	4399	46	4456	46	4489
48	4448	48	4595	48	4655	48	4688
50	4647	50	4794	50	4854	50	4887
52	4848	52	4995	52	5055	52	5087
54	5053	54	5198	54	5257	54	5289
56	5260	56	5402	56	5460	56	5491
58	5470	58	5607	58	5663	58	5694
60	5683	60	5814	60	5869	60	5898
62	5898	62	6023	62	6075	62	6103
64	6116	64	6233	64	6282	64	6308
66	6336	66	6444	66	6489	66	6514
68	6557	68	6656	68	6697	68	6721
70	6779	70	6869	70	6906	70	6928
72	7001	72	7082	72	7115	72	7135
74	7226	74	7295	74	7325	74	7342
76	7449	76	7508	76	7534	76	7548
78	7671	78	7721	78	7743	78	7755
80	7892	80	7934	80	7952	80	7963
82	8112	82	8146	82	8160	82	8169
84	8330	84	8357	84	8369	84	8375
86	8547	86	8567	86	8576	86	8580
88	8763	88	8777	88	8782	88	8785
90	8974	90	8984	90	8988	90	8990
92	9183	92	9190	92	9192	92	9194
94	9391	94	9394	94	9396	94	9397
96	9596	96	9598	96	9599	96	9599
98	9800	98	9800	98	9800	98	9800

Т а б л и ц а IV

Коэффициенты первых 15 смещенных полиномов Лежандра $P_n^*(x)$
(см. стр. 295)

(Подчеркнутые числа отрицательны; последовательность степеней: от низших к высшим, например $P_3^*(x) = 1 - 12x + 30x^2 - 20x^3$.)

$n = 0:$	1
$n = 1:$	1, <u>2</u>
$n = 2:$	1, <u>6</u> , 6
$n = 3:$	1, <u>12</u> , 30, <u>20</u>
$n = 4:$	1, <u>20</u> , 90, <u>140</u> , 70
$n = 5:$	1, <u>30</u> , 210, <u>560</u> , 630, <u>252</u>
$n = 6:$	1, <u>42</u> , 420, <u>1680</u> , 3150, <u>2772</u> , 924
$n = 7:$	1, <u>56</u> , 756, <u>4200</u> , 11550, <u>16632</u> , 12012, <u>3432</u>
$n = 8:$	1, <u>72</u> , 1260, <u>9240</u> , 34650, <u>72072</u> , 84084, <u>51480</u> , 12870
$n = 9:$	1, <u>90</u> , 1980, <u>18480</u> , 90090, <u>252252</u> , 420420, <u>411840</u> , 218790, <u>48620</u>
$n = 10:$	1, <u>110</u> , 2970, <u>34320</u> , 210210, <u>756756</u> , 1681680, <u>2333760</u> , 1969110, <u>923780</u> , 184756
$n = 11:$	1, <u>132</u> , 4290, <u>60060</u> , 450450, <u>2018016</u> , 5717712, <u>10501920</u> , 12471030, <u>9237800</u> , 3879876, <u>705432</u>
$n = 12:$	1, <u>156</u> , 6006, <u>100100</u> , 900900, <u>4900896</u> , 17153136, <u>39907296</u> , 62355150, <u>64664600</u> , 42678636, <u>16224936</u> , 2704156
$n = 13:$	1, <u>182</u> , 8190, <u>160160</u> , 1701700, <u>11027016</u> , 46558512, <u>133024320</u> , 261891630, <u>355655300</u> , 327202876, <u>194699232</u> , 97603900, <u>10400600</u>
$n = 14:$	1, <u>210</u> , 10920, <u>247520</u> , 3063060, <u>23279256</u> , 116396280, <u>399072960</u> , 960269310, <u>1636014380</u> , 1963217256, <u>1622493600</u> , 878850700, <u>280816200</u> , 40116600
$n = 15:$	1, <u>240</u> , 14280, <u>371280</u> , 5290740, <u>46558512</u> , 271591320, <u>1097450640</u> , 3155170590, <u>6544057520</u> , 9816086280, <u>10546208400</u> , 7909656300, <u>3931426800</u> , 1163381400, <u>155117520</u>

Т а б л и ц а V

Первые 12 полиномов Чебышева $T_n(x)$ (см. стр. 449)

$$T_0(x) = 1 \text{ (но } T_0 = \frac{1}{2} \text{ ; см. стр. 451)}$$

$$T_1(x) = x$$

$$T_2(x) = 2x^2 - 1$$

$$T_3(x) = 4x^3 - 3x$$

$$T_4(x) = 8x^4 - 8x^2 + 1$$

$$T_5(x) = 16x^5 - 20x^3 + 5x$$

$$T_6(x) = 32x^6 - 48x^4 + 18x^2 - 1$$

$$T_7(x) = 64x^7 - 112x^5 + 56x^3 - 7x$$

$$T_8(x) = 128x^8 - 256x^6 + 160x^4 - 32x^2 + 1$$

$$T_9(x) = 256x^9 - 576x^7 + 432x^5 - 120x^3 + 9x$$

$$T_{10}(x) = 512x^{10} - 1280x^8 + 1120x^6 - 400x^4 + 50x^2 - 1$$

$$T_{11}(x) = 1024x^{11} - 2816x^9 + 2816x^7 - 1232x^5 + 220x^3 - 11x$$

$$T_{12}(x) = 2048x^{12} - 6144x^{10} + 6912x^8 - 3584x^6 + 840x^4 - 72x^2 + 1$$

Т а б л и ц а VI

Первые 12 степеней $\frac{1}{2}(2x)^n$, выраженные как линейные комбинации чебышевских полиномов $T_k(x)$ ($T_0 = \frac{1}{2}$)

	T_0	T_2	T_4	T_6	T_8	T_{10}	T_{12}
$n=0:$	1						
$n=2:$	2	1					
$n=4:$	6	4	1				
$n=6:$	20	15	6	1			
$n=8:$	70	56	28	8	1		
$n=10:$	252	210	120	45	10	1	
$n=12:$	924	792	495	220	66	12	1

	T_1	T_3	T_5	T_7	T_9	T_{11}
$n=1:$	1					
$n=3:$	3	1				
$n=5:$	10	5	1			
$n=7:$	35	21	7	1		
$n=9:$	126	84	36	9	1	
$n=11:$	462	330	165	55	11	1

Например: $32x^6 = 20T_0 + 15T_2 + 6T_4 + T_6$, $256x^9 = 126T_1 + 84T_3 + 36T_5 + 9T_7 + T_9$.

Т а б л и ц а VII

Коэффициенты первых 12 смещенных полиномов Чебышева $T_n^*(x)$
(см. стр. 450)

(Подчеркнутые числа отрицательны; последовательность — от низших степеней к высшим, например $T_3^*(x) = -1 + 18x - 48x^2 + 32x^3$.)

$T_0^*(x) = 1$ (но $T_0^* = \frac{1}{2}$; см. стр. 451)

$n = 1:$	1, 2
$n = 2:$	<u>1</u> , 8, 8
$n = 3:$	1, <u>18</u> , 48, 32
$n = 4:$	<u>1</u> , 32, <u>160</u> , 256, 128
$n = 5:$	1, <u>50</u> , 400, <u>1120</u> , 1280, 512
$n = 6:$	<u>1</u> , 72, <u>840</u> , 3584, <u>6912</u> , <u>6144</u> , 2048
$n = 7:$	1, <u>98</u> , 1568, <u>9408</u> , 26880, <u>39424</u> , 28672, 8192
$n = 8:$	<u>1</u> , 128, <u>2688</u> , 21504, <u>84480</u> , 180224, <u>212992</u> , 131072, 32768
$n = 9:$	1, <u>162</u> , 4320, <u>44352</u> , 228096, <u>658944</u> , 1118208, <u>1105920</u> , <u>589824</u> , 131072
$n = 10:$	<u>1</u> , 200, <u>6600</u> , <u>84480</u> , <u>549120</u> , <u>2050048</u> , <u>4659200</u> , <u>6553600</u> , <u>5570560</u> , <u>2621440</u> , <u>524288</u>
$n = 11:$	1, 242, <u>9680</u> , 151008, <u>1208064</u> , 5637632, <u>16400384</u> , 30638080, <u>36765696</u> , <u>27394048</u> , <u>11534336</u> , 2097152
$n = 12:$	1, <u>288</u> , 13728, <u>256256</u> , 2471040, 14057472, 50692096, <u>120324096</u> , <u>190513152</u> , <u>199229440</u> , <u>132120576</u> , <u>50331648</u> , 8388608

Т а б л и ц а VIII

Первые 12 степеней $\frac{1}{2}(4x)^n$, выраженные как линейные комбинации смещенных полиномов Чебышева $T_k^*(x)$ ($T_0^* = \frac{1}{2}$)
(см. стр. 455)

	T_0^*	T_1^*	T_2^*	T_3^*	T_4^*	T_5^*	T_6^*
$n = 0:$	1						
$n = 1:$	2	1					
$n = 2:$	6	4	1				
$n = 3:$	20	15	6	1			
$n = 4:$	70	56	28	8	1		
$n = 5:$	252	210	120	45	10	1	
$n = 6:$	924	792	495	220	66	12	1
$n = 7:$	3432	3003	2002	1001	364	91	14
$n = 8:$	12870	11440	8008	4368	1820	560	120
$n = 9:$	48620	43758	31824	18564	8568	3060	816
$n = 10:$	184756	167960	125970	77520	38760	15504	4845
$n = 11:$	705432	646646	497420	319770	170544	74613	26334
$n = 12:$	2704156	2496144	1961256	1307504	735471	346104	134596

Продолжение табл. VIII

	T_7^*	T_8^*	T_9^*	T_{10}^*	T_{11}^*	T_{12}^*
$n=7$:	1					
$n=8$:	16	1				
$n=9$:	153	18	1			
$n=10$:	1140	190	20	1		
$n=11$:	7315	1540	231	22	1	
$n=12$:	42504	10626	2024	276	24	1

Например: $32\,768x^8 = 12\,870T_0^* + 11\,440T_1^* + 8008T_2^* + 4368T_3^* + 1820T_4^* + 560T_5^* + 120T_6^* + 16T_7^* + T_8^*$.

Т а б л и ц а IX

Коэффициенты первых 13 смещенных полиномов Чебышева
второго рода $U_n^*(x)$ (см. стр. 297)

(Подчеркнутые числа отрицательны; последовательность — от низших к высшим степеням, например $U_3^*(x) = -4 + 40x - 96x^2 + 64x^3$.)

$n=0$:	1
$n=1$:	<u>2</u> , 4
$n=2$:	3, <u>16</u> , 16
$n=3$:	<u>4</u> , 40, <u>96</u> , 64
$n=4$:	5, <u>80</u> , 336, <u>512</u> , 256
$n=5$:	<u>6</u> , 140, <u>896</u> , 2304, <u>2560</u> , 1024
$n=6$:	7, <u>224</u> , 2016, <u>7680</u> , 14080, <u>12288</u> , 4096
$n=7$:	<u>8</u> , 336, <u>4032</u> , 21120, <u>56320</u> , 79872, <u>57344</u> , 16384
$n=8$:	9, <u>480</u> , 7392, <u>50688</u> , 183040, <u>372736</u> , 430080, <u>262144</u> , 65536
$n=9$:	<u>10</u> , 660, <u>12672</u> , 109824, <u>512512</u> , 1397760, <u>2293760</u> , <u>2228224</u> , <u>1179648</u> , 262144
$n=10$:	11, <u>880</u> , 20592, <u>219648</u> , 1281280, <u>4472832</u> , 9748480, <u>13369344</u> , 11206656, <u>5242880</u> , 1048576
$n=11$:	<u>12</u> , 1144, <u>32032</u> , 411840, <u>2928640</u> , 12673024, <u>35094528</u> , 63504384, <u>74711040</u> , <u>55050240</u> , <u>23068672</u> , 4194304
$n=12$:	13, <u>1456</u> , 48048, <u>732160</u> , 6223360, <u>32587776</u> , 111132672, <u>254017536</u> , 392232960, <u>403701760</u> , 265289728, <u>100663296</u> , 16777216
$n=13$:	<u>14</u> , 1820, <u>69888</u> , 1244672, <u>12446720</u> , 77395968, <u>317521920</u> , 889061376, <u>1725825024</u> , 2321285120, <u>2122317824</u> , 1258291200, <u>436207616</u> , 67108864

Т а б л и ц а X

Коэффициенты первых 11 полиномов Лагерра $L_n(x)$ (см. стр. 305)(Подчеркнутые числа отрицательны; последовательность — от низших степеней к высшим, например $L_3(x) = 6 - 18x + 9x^2 - x^3$.)

$n=0$:	1
$n=1$:	1, <u>1</u>
$n=2$:	2, <u>4</u> , 1
$n=3$:	6, <u>18</u> , 9, <u>1</u>
$n=4$:	24, <u>96</u> , 72, <u>16</u> , 1
$n=5$:	120, <u>600</u> , 600, <u>200</u> , 25, <u>1</u>
$n=6$:	720, <u>4320</u> , 5400, <u>2400</u> , 450, <u>36</u> , 1
$n=7$:	5040, <u>35280</u> , 52920, <u>29400</u> , 7350, <u>882</u> , 49, <u>1</u>
$n=8$:	40320, <u>322560</u> , 564480, <u>376320</u> , 117600, <u>18816</u> , 1568, <u>64</u> , 1
$n=9$:	362880, <u>3265920</u> , 6531840, <u>5080320</u> , 1905120, <u>381024</u> , 42336, <u>2592</u> , 81, <u>1</u>
$n=10$:	3628800, <u>36288000</u> , 81648000, <u>72576000</u> , 31752000, <u>7620480</u> , 1058400, <u>86400</u> , 4050, <u>100</u> , 1
$n=11$:	39916800, <u>439084800</u> , 1097712000, <u>1097712000</u> , 548856000, <u>153679680</u> , 25613280, <u>2613600</u> , 163350, <u>6050</u> , 121, <u>1</u>

Значения нормированных функций Лагерра

Таблица значений нормированных функций Лагерра (от удвоенного аргумента)

$$\varphi_n(x) = \frac{e^{-x}}{n!} L_n(2x) \quad (n \leq 20; \text{ см. стр. 305}).$$

(Вычислена под наблюдением Национального Бюро Стандартов (Вашингтон) (в оригинале с 10 десятичными знаками). Подчеркнутые числа отрицательны.

Рекуррентная формула:

$$\varphi_{n+1}(x) = \frac{1}{n+1} [(2n+1-2x)\varphi_n(x) - n\varphi_{n-1}(x)].$$

Продолжение табл. X

n	$x = 0$	0.1	0.2	0.3	0.4	0.5	1	1.5	2
0	1	.90484	.81873	.74082	.67032	.60653	.36788	.22313	.13534
1	1	72387	49124	29633	13406	0	36788	44626	40601
2	1	56100	22924	01482	18769	30327	36788	11157	13534
3	1	41502	02402	21928	35214	40435	12263	22313	31578
4	1	28478	13231	33974	40505	37908	12263	30680	13534
5	1	16920	24678	39534	38257	28305	26978	18966	11729
6	1	06725	32572	40213	31283	15584	30248	00279	24962
7	1	02207	37478	37349	21730	02455	24409	16655	22040
8	1	09967	39896	32042	11198	09340	13197	24739	08464
9	1	16643	40272	25188	00841	18787	00298	23678	07366
10	1	22318	39000	17508	08548	25410	11370	15620	18666
11	1	27072	36425	09572	16461	29122	19910	04034	22152
12	1	30978	32851	01818	22618	30097	24420	07594	17963
13	1	34108	28541	05423	26909	28682	24827	16576	08569
14	1	36527	23723	11915	29356	25319	21657	21364	02602
15	1	38297	18594	17497	30074	20493	15812	21560	12334
16	1	39478	13321	22074	29244	14687	08355	17701	18375
17	1	40125	08044	25603	27086	08359	00354	10946	19737
18	1	40290	02882	28082	23844	01918	07242	02741	16637
19	1	40023	02070	29544	19769	04284	13676	05464	10197
20	1	39368	06732	30047	15107	09963	18421	12440	02041

n	$x = 2.5$	3	3.5	4	4.5	5	5.5	6	6.5
0	.08208	.04979	.03020	.01832	.01111	.00674	.00409	.00248	.00150
1	32834	24894	18118	12821	08887	06064	04087	02727	01804
2	28730	34851	34727	31137	26106	20888	16143	12146	08945
3	21889	04979	11072	22589	28883	30770	29561	26523	22652
4	10603	24894	26045	17705	05138	07412	17454	24044	27269
5	25994	18919	01560	14530	23107	23134	16667	06792	03693
6	17158	04979	20664	22019	11984	02321	14545	21169	21493
7	02671	21195	19050	03274	12958	20823	18442	08846	03166
8	18352	19488	00968	16403	20204	10984	03506	15205	19598
9	22095	04979	15857	19313	06441	09967	18730	16310	05896
10	14416	11067	19900	06481	11743	18856	11829	02268	14100
11	00883	19617	10912	09897	18666	09795	06274	16683	15615
12	11891	17647	03693	18313	11013	06673	17117	13214	00087
13	19109	07683	15185	14811	03676	16741	12642	02186	14333
14	18986	04862	18264	03097	14952	14132	01446	14612	14414
15	12543	14626	12615	09489	16505	02276	13535	14520	01997
16	02583	18294	01799	16543	08677	10262	15563	03544	11266
17	07551	15291	09121	15398	03284	16026	07402	09288	15134
18	15024	07356	15887	07472	12939	12566	04829	15215	07857
19	18150	02483	16444	03182	15957	02675	13621	11221	04413
20	16583	11085	11218	12031	11643	08059	14481	00694	13201

Продолжение табл. X

<i>n</i>	<i>x</i> = 7	7.5	8	8.5	9	9.5	10	11	12
0	.00091	.00055	.00034	.00020	.00012	.00007	.00005	.00002	.00001
1	01185	00774	00503	00326	00210	00135	00086	00035	00014
2	06474	04618	03254	02269	01567	01074	00731	00332	00148
3	18633	14878	11596	08858	06652	04923	03597	01860	00928
4	27752	26292	23650	20443	17117	13963	11143	06726	03835
5	12845	19648	23834	25622	25489	23987	21636	15999	10761
6	16704	08811	00153	08587	15473	20347	23169	23727	20120
7	13397	19359	20363	17055	10796	03120	04623	16792	22394
8	16290	07710	02679	11777	17587	19364	17383	06068	07588
9	06478	15495	18398	15160	07642	01530	09904	18297	14004
10	17900	13137	03108	07567	15064	17427	14654	00028	13831
11	05502	06921	15313	16534	11056	01778	07671	16631	08959
12	12282	16656	11782	01330	09203	15383	15351	01360	13425
13	15471	06424	05978	14443	15160	08741	01177	15038	09302
14	02961	09960	15638	11552	01201	09289	14843	06634	10473
15	11478	15291	07973	04239	13269	14351	07807	10940	12173
16	14972	05954	07186	14539	11907	02055	08548	12373	04492
17	05930	08088	14689	09694	01982	11815	13884	02291	13835
18	07222	14610	08719	04037	13118	12442	03497	13340	04212
19	14360	09254	04280	13433	11240	00595	10024	08361	10225
20	11090	02774	13205	10942	00660	11226	12846	05566	11670

<i>n</i>	<i>x</i> = 13	14	15	16	17	18	19	20	21
0	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000
1	00006	00002	00001	00000	00000	00000	00000	00000	00000
2	00065	00028	00012	00005	00002	00001	00000	00000	00000
3	00450	00213	00099	00045	00020	00009	00004	00002	00001
4	02090	01099	00561	00279	00136	00065	00031	00014	00006
5	06747	04006	02276	01247	00663	00343	00174	00086	00042
6	15126	10433	06740	04132	02428	01376	00757	00405	00212
7	22308	18924	14417	10148	06715	04228	02554	01490	00843
8	17438	21622	21135	17948	13825	09895	06681	04302	02661
9	02391	09606	17713	20893	20144	17130	13319	09668	06641
10	17368	10815	00463	11008	17774	20216	19294	16432	12881
11	05721	15615	15724	07986	02693	11995	17709	19594	18553
12	14490	03407	09596	16080	13825	05537	04450	12695	17568
13	06395	15200	10823	01287	12056	15758	11896	03439	05848
14	13912	02078	11230	14472	06809	04988	13480	14981	10048
15	03187	14048	09353	04095	13522	12379	03016	07777	14166
16	14039	04582	09944	13311	03848	08545	13956	09670	05110
17	02781	11874	10558	03071	12953	10143	01267	11301	10627
18	11868	08945	06459	13084	04354	08634	12970	05994	08959
19	09506	07012	12382	00533	11584	10064	01883	11653	07710
20	05096	12355	00564	12243	07032	06692	12416	05112	09667

Т а б л и ц а X I

Преобразование гармонического ряда в ряд, расположенный по полиномам Чебышева

Ряд

$$f(x)=\sum_{k=0}^N a_k \cos (2k+1) \frac{\pi}{2} x+\sum_{k=1}^N b_k \sin k \pi x$$

преобразуется в ряд

$$f(x)=\sum_{i=0}^{\infty} c_i T_i(x)$$

(если пренебречь всеми c_i , для которых $i>13$) путем умножения чисел a_k на последовательные столбцы первой матрицы и чисел b_k на последовательные столбцы второй матрицы (см. стр. 354). (Подчеркнутые числа отрицательны; $N\leq 11$.)

k	c_0	c_2	c_4	c_6	c_8	c_{10}	c_{12}
0	.472001	.499403	.027992	.000597	.000007	.000000	.000000
1	265857	292636	740869	205626	024909	001735	000079
2	204268	300940	138934	691873	418338	107419	016207
3	171971	279881	032140	402029	451227	560378	242595
4	151323	259041	097117	211630	465982	121520	567021
5	136676	241238	125502	101019	361305	347427	181560
6	125594	226275	138516	034324	262471	374524	116288
7	116832	213613	144229	007805	185551	337894	256291
8	109679	202772	146188	035480	127870	286247	303192
9	103697	193378	146117	054225	084651	235836	304120
10	098598	185148	144919	067216	051946	191405	284458
11	094184	177867	143093	076366	026880	153695	256880

k	c_1	c_3	c_5	c_7	c_9	c_{11}	c_{13}
1	.569231	.666917	.104282	.006841	.000250	.000006	.000000
2	424765	058224	745649	315042	058248	006296	000453
3	353450	167800	294175	589723	503360	170328	033705
4	309062	193132	097432	459239	290219	582676	318085
5	278050	197132	003055	301849	426211	040189	513869
6	254818	194304	047027	189771	385862	240016	291674
7	236577	189152	075527	113651	313247	329561	010673
8	221764	183315	092486	061327	244663	334392	162549
9	209424	177443	102825	024535	187474	306725	247246
10	198938	171797	109159	001931	141479	269136	278343
11	189884	166476	112969	021356	104801	230904	280138

Т а б л и ц а XII

Подходящие кривые для равноотстоящих данных

Для заданных $2n + 1$ равноотстоящих ординат y_k [$k = -n, -(n - 1), \dots, (n - 1), n$] подходящим бесконечным разложением является

$$f(x) = \sum_{i=0}^{\infty} c_i T_i(x).$$

Ряд может быть усечен при надлежаще выбранном $i = \nu$ (см. стр. 354). Последние две ординаты приведены к нулю (см. стр. 337). Остальные $2n - 1$ ординат разбиты на две группы:

$$u_k = y_k + y_{-k} \quad (k = 0, 1, 2, \dots, n - 1)$$

и

$$v_k = y_k - y_{-k} \quad (k = 1, 2, \dots, n - 1).$$

Четные коэффициенты c_{2m} вычислены путем умножения u_k на последовательные столбцы одной из таблиц первой группы соответственно числу заданных точек (таблица рассчитана для случаев от 5 до 25 заданных точек, т. е. от $n = 2$ до $n = 12$; отрицательные числа подчеркнуты).

Нечетные коэффициенты c_{2m+1} вычислены аналогично путем умножения v_k на последовательные столбцы одной из таблиц второй группы.

Четные c_i

k	c_0	c_2	c_4	c_6	c_8	c_{10}	c_{12}
$n = 2$							
0	.051536	.198010	.192215	.051556	.006229	.000434	.000020
1	260872	<u>073103</u>	<u>252040</u>	<u>072489</u>	<u>008804</u>	<u>000614</u>	<u>000028</u>
$n = 3$							
0	.068402	.081850	.104988	.149683	.073876	.018192	.002715
1	077288	<u>231039</u>	048188	<u>199554</u>	<u>120762</u>	<u>031009</u>	<u>004679</u>
2	201331	<u>064468</u>	<u>265447</u>	<u>046870</u>	<u>061421</u>	<u>017325</u>	<u>002675</u>
$n = 4$							
0	.029805	.096373	.074723	.062008	.111810	.083692	.032360
1	103761	<u>107491</u>	098060	<u>046475</u>	141858	<u>139541</u>	057575
2	063926	<u>139227</u>	107141	<u>229621</u>	<u>001412</u>	079766	<u>040006</u>
3	170194	<u>116096</u>	<u>197455</u>	<u>150828</u>	047702	<u>029202</u>	<u>019484</u>
$n = 5$							
0	.038976	.051194	.069490	.070770	.042850	.079105	.082590
1	049960	<u>145764</u>	077724	<u>031293</u>	038520	<u>088787</u>	<u>136363</u>
2	087062	<u>063696</u>	004228	<u>091897</u>	188491	036562	<u>073506</u>
3	055556	<u>086734</u>	155161	140391	<u>135871</u>	091975	<u>020528</u>
4	150219	<u>137968</u>	<u>134730</u>	<u>183270</u>	<u>022172</u>	<u>061958</u>	<u>000980</u>
	c_0	c_2	c_4	c_6	c_8	c_{10}	c_{12}

Продолжение табл. XII

Четные c_i

k	c_0	c_2	c_4	c_6	c_8	c_{10}	c_{12}
$n = 6$							
0	.021091	.062765	.047450	.050557	.065817	.036969	.053695
1	065054	<u>081523</u>	095971	<u>062839</u>	001731	<u>022276</u>	047351
2	043738	<u>109942</u>	010618	<u>056330</u>	<u>073360</u>	<u>146535</u>	<u>063561</u>
3	076558	<u>033860</u>	045192	116235	<u>096555</u>	<u>108443</u>	<u>114892</u>
4	049724	<u>054366</u>	<u>162047</u>	053758	<u>176085</u>	<u>006155</u>	088080
5	135991	<u>147385</u>	<u>086438</u>	<u>181196</u>	<u>078732</u>	<u>043817</u>	<u>038201</u>
$n = 7$							
0	.027049	.037636	.05066	.045786	.037667	.058439	.037718
1	037099	<u>104213</u>	066741	<u>059319</u>	053798	<u>014038</u>	002342
2	057015	<u>062020</u>	049309	<u>015222</u>	081722	063230	107341
3	039324	<u>083701</u>	033169	089660	<u>072997</u>	061509	<u>078444</u>
4	069167	<u>013397</u>	<u>067635</u>	096508	<u>007112</u>	150629	029117
5	045377	<u>033142</u>	<u>153880</u>	008840	<u>161891</u>	<u>073419</u>	079610
6	125186	<u>150928</u>	<u>050308</u>	<u>165515</u>	<u>113337</u>	007891	<u>052879</u>
$n = 8$							
0	.016366	.046282	.035231	.040550	.044555	.030016	.049022
1	047298	<u>064460</u>	079634	<u>057019</u>	033449	<u>044969</u>	013504
2	033458	<u>085966</u>	033788	<u>006173</u>	025525	<u>088291</u>	<u>061550</u>
3	051398	<u>046093</u>	013834	<u>059517</u>	<u>096981</u>	042031	<u>031191</u>
4	035991	<u>064276</u>	058885	089520	<u>032817</u>	018363	119100
5	063597	<u>000980</u>	<u>075649</u>	065946	<u>048602</u>	<u>125873</u>	060961
6	041982	018552	<u>140763</u>	050034	127054	<u>119952</u>	<u>032144</u>
7	116615	<u>151489</u>	<u>023181</u>	<u>145570</u>	<u>131003</u>	<u>027599</u>	<u>045997</u>
$n = 9$							
0	.020641	.029875	.039438	.034074	.032501	.042584	.026731
1	029584	<u>080574</u>	055967	<u>058507</u>	050576	<u>016780</u>	034104
2	042566	<u>053695</u>	053968	<u>013507</u>	022049	<u>029715</u>	083680
3	030690	<u>071033</u>	006341	<u>030837</u>	<u>061859</u>	089196	<u>024389</u>
4	047188	<u>033442</u>	010528	076840	<u>074948</u>	010331	<u>034580</u>
5	033365	<u>049612</u>	<u>072851</u>	074385	<u>008558</u>	<u>056361</u>	087685
6	059201	<u>011327</u>	<u>076189</u>	036770	075088	<u>078357</u>	110501
7	039239	008150	<u>126688</u>	075785	088391	<u>137623</u>	<u>020381</u>
8	109600	<u>150469</u>	<u>002571</u>	<u>125318</u>	<u>137505</u>	<u>056487</u>	028018
	c_0	c_2	c_4	c_6	c_8	c_{10}	c_{12}

Продолжение табл. XII

Четные c_i

k	c_0	c_2	c_4	c_6	c_8	c_{10}	c_{12}
$n = 10$							
0	.013392	.036556	.028188	.033378	.033483	.026533	.039264
1	037102	<u>053135</u>	066292	<u>049638</u>	041173	<u>045205</u>	008011
2	027181	<u>069712</u>	037353	<u>026308</u>	000860	<u>038969</u>	032977
3	039031	<u>044121</u>	031191	<u>020473</u>	054546	054572	<u>074808</u>
4	028497	<u>058906</u>	014235	050479	<u>064660</u>	052097	<u>034002</u>
5	043881	<u>023409</u>	<u>026439</u>	077959	<u>041281</u>	051607	052337
6	031230	<u>038319</u>	<u>079522</u>	054432	<u>038873</u>	<u>060942</u>	036059
7	055617	<u>018929</u>	<u>072962</u>	012561	082404	<u>031420</u>	121822
8	036965	000518	<u>113214</u>	091063	052829	<u>136250</u>	<u>062863</u>
9	103719	<u>148599</u>	<u>013294</u>	<u>106345</u>	<u>137101</u>	<u>077852</u>	006458
$n = 11$							
0	.016656	.024817	.032213	.027288	.028078	.032821	.022764
1	024638	<u>065480</u>	047686	<u>052993</u>	045107	<u>027805</u>	039226
2	033986	<u>046377</u>	050633	<u>023412</u>	004566	<u>006902</u>	<u>047221</u>
3	025270	<u>060263</u>	019820	<u>001567</u>	034004	064401	<u>045676</u>
4	036261	<u>035906</u>	012886	042138	<u>063299</u>	042768	<u>029841</u>
5	026706	<u>049020</u>	028905	057472	<u>050171</u>	009798	<u>069210</u>
6	041192	<u>015398</u>	<u>036391</u>	070649	<u>009167</u>	071840	036492
7	029451	<u>029462</u>	<u>081739</u>	034481	<u>057001</u>	<u>047272</u>	007926
8	052619	<u>024610</u>	<u>067949</u>	006292	078671	<u>006239</u>	<u>109018</u>
9	035040	005211	<u>100895</u>	<u>099390</u>	022613	124033	<u>091632</u>
10	098695	146280	<u>025659</u>	<u>089188</u>	<u>132603</u>	<u>092591</u>	<u>014586</u>
$n = 12$							
0	.011344	.030160	.023566	.028196	.026858	.023682	.031571
1	030490	<u>045167</u>	056270	<u>043336</u>	041143	<u>041494</u>	018535
2	022928	<u>058339</u>	036051	<u>032318</u>	013014	011452	012403
3	031550	<u>040077</u>	035516	<u>000623</u>	024575	038615	<u>065383</u>
4	023704	<u>052155</u>	004974	021530	<u>050563</u>	060969	<u>021701</u>
5	034014	<u>028939</u>	001160	053580	<u>056613</u>	016386	<u>012406</u>
6	025208	<u>040898</u>	<u>039004</u>	056505	<u>029756</u>	022671	075513
7	038950	<u>008945</u>	<u>042256</u>	059622	016473	<u>074394</u>	007693
8	027940	<u>022402</u>	<u>081269</u>	016535	065292	<u>026587</u>	035796
9	050064	<u>028913</u>	<u>062210</u>	020439	069251	<u>033045</u>	<u>084605</u>
10	033386	009591	<u>189860</u>	<u>103146</u>	002024	106515	<u>107898</u>
11	094339	143741	<u>035405</u>	<u>073931</u>	<u>125799</u>	<u>102028</u>	<u>033242</u>
	c_0	c_2	c_4	c_6	c_8	c_{10}	c_{12}

Продолжение табл. XII

Нечетные c_i

k	c_1	c_3	c_5	c_7	c_9	c_{11}
$n = 2$						
1	.284615	<u>.333458</u>	.052141	<u>.003420</u>	.000125	<u>.000003</u>
$n = 3$						
1	.041704	<u>.209330</u>	.245354	<u>.092920</u>	.016887	<u>.001819</u>
2	286942	<u>175714</u>	<u>185147</u>	<u>088970</u>	<u>016743</u>	<u>001816</u>
$n = 4$						
1	.056917	<u>.102788</u>	.152844	<u>.184219</u>	.103589	<u>.031685</u>
2	053945	<u>208679</u>	099615	<u>145721</u>	125778	<u>042581</u>
3	269300	<u>073676</u>	<u>219981</u>	<u>026698</u>	<u>074465</u>	<u>028537</u>
$n = 5$						
1	.017020	<u>.080262</u>	.109589	<u>.118914</u>	.140971	<u>.102094</u>
2	075576	<u>116690</u>	123542	<u>056363</u>	107482	<u>130114</u>
3	057871	<u>176472</u>	014706	<u>192413</u>	<u>010771</u>	<u>090069</u>
4	251275	<u>012704</u>	<u>196981</u>	<u>107038</u>	<u>050578</u>	<u>037296</u>
$n = 6$						
1	.023596	<u>.047462</u>	.081095	<u>.103198</u>	.098694	<u>.110050</u>
2	025328	<u>105242</u>	109055	<u>069177</u>	028072	<u>077392</u>
3	082305	<u>106264</u>	065901	<u>046839</u>	154887	<u>035085</u>
4	058729	<u>144187</u>	078069	154339	<u>095037</u>	<u>088995</u>
5	235434	<u>025098</u>	<u>162282</u>	<u>144824</u>	<u>001900</u>	<u>059971</u>
$n = 7$						
1	.009284	<u>.041395</u>	.059088	<u>.075735</u>	.093361	<u>.086213</u>
2	035218	<u>065979</u>	096903	<u>088815</u>	037615	<u>007724</u>
3	029738	<u>108508</u>	075977	<u>001815</u>	057667	<u>119632</u>
4	084336	<u>090308</u>	018405	090593	<u>107540</u>	<u>076603</u>
5	058302	<u>117106</u>	109225	098261	<u>143562</u>	<u>024587</u>
6	221847	<u>049496</u>	<u>128786</u>	<u>156808</u>	<u>048326</u>	<u>047744</u>
$n = 8$						
1	.012775	<u>.027172</u>	.048205	<u>.062395</u>	.069469	<u>.082970</u>
2	014824	<u>061249</u>	075895	<u>079106</u>	068921	<u>020267</u>
3	041416	<u>070810</u>	084554	<u>039744</u>	036860	<u>061658</u>
4	032157	<u>103342</u>	039985	<u>049336</u>	<u>077009</u>	<u>067509</u>
5	084285	<u>074452</u>	014588	097211	<u>042814</u>	<u>125327</u>
6	057311	<u>095269</u>	<u>122275</u>	047098	<u>150825</u>	<u>041311</u>
7	210175	<u>065752</u>	<u>099653</u>	<u>155059</u>	<u>081594</u>	<u>021382</u>
	c_1	c_3	c_5	c_7	c_9	c_{11}

Нечетные c_i

k	c_1	c_3	c_5	c_7	c_9	c_{11}
$n = 9$						
1	.005862	.025072	.036592	.049810	.061780	.063935
2	020347	041595	067417	072989	061726	051666
3	018311	069600	071071	051153	012115	052059
4	044822	069289	063602	005537	075101	059665
5	033467	095175	009657	071980	058788	003902
6	083256	060284	035879	086055	009673	123219
7	056078	077753	125432	006334	136128	087954
8	200067	076837	075161	146494	102908	007125
$n = 10$						
1	.007971	.017579	.031591	.041247	.049078	.059205
2	009769	039695	053127	064193	065347	047586
3	025078	048481	069461	056367	022419	007064
4	020577	071807	057395	018085	032006	078029
5	046668	064929	042270	037015	080355	023335
6	034123	086290	031656	077416	028284	038406
7	081767	048083	048853	068189	044280	095626
8	054759	063627	123278	024281	112049	113930
9	191231	084510	054780	134870	115130	032598
$n = 11$						
1	.004044	.016762	.024808	.034725	.043136	.047442
2	013231	028467	048131	056080	057319	056309
3	012471	047790	056721	056712	038800	001420
4	028127	050967	062849	031079	015812	043909
5	022074	070744	041325	009523	054054	066720
6	047592	059447	023518	055217	066528	016345
7	034379	077647	030769	072744	001351	056299
8	080075	037714	056167	049132	063510	060290
9	053432	052131	118362	046511	085511	123727
10	183434	089860	037839	122190	120867	053386
$n = 12$						
1	.005433	.012301	.022202	.029152	.035916	.043349
2	006939	027708	038677	049996	054521	049668
3	016834	034884	054216	053085	040515	022764
	c_1	c_3	c_5	c_7	c_9	c_{11}

Продолжение табл. XII

<i>k</i>	<i>c</i> ₁	<i>c</i> ₃	<i>c</i> ₅	<i>c</i> ₇	<i>c</i> ₉	<i>c</i> ₁₁
4	014386	051816	053029	039767	006439	036285
5	030123	050903	052570	006559	041103	052154
6	023066	067981	025811	029066	058229	038688
7	047949	053686	008081	063354	044854	045589
8	034381	069626	042940	062861	025028	056327
9	078315	028931	059704	031372	071537	026568
10	052134	042684	112116	062235	059959	122455
11	176496	093587	023715	109503	122242	069418
	<i>c</i> ₁	<i>c</i> ₃	<i>c</i> ₅	<i>c</i> ₇	<i>c</i> ₉	<i>c</i> ₁₁

Таблица XIII
Нули и веса
гауссовой квадратуры
(см. стр. 408 и 412)

<i>n</i>	$\pm x_i$	<i>w</i> _{<i>i</i>}
2	0.5773502692	1
3	0 0.7745966692	0.8888888889 0.5555555556
4	0.3399810435 0.8611363116	0.6521451549 0.3478548451
5	0 0.5384693101 0.9061798459	0.5688888889 0.4786286705 0.2369268851
6	0.2386191861 0.6612093865 0.9324695142	0.4679139346 0.3607615730 0.1713244924
7	0 0.4058451514 0.7415311856 0.9491079123	0.4179591837 0.3818300505 0.2797053915 0.1294849662
8	0.1834346425 0.5255324099 0.7966664774 0.9602898565	0.3626837834 0.3137066459 0.2238103445 0.1012285363
9	0 0.3242534234 0.6133714327 0.8360311073 0.9681602395	0.3302393550 0.3123470770 0.2606106964 0.1806481607 0.0812743884

Таблица XIV
Гауссова квадратура
с округленными узлами
(см. стр. 408 и 412)

<i>n</i>	$\pm x_i$	<i>w</i> _{<i>i</i>}
2	0.58	1
3	0 0.77	0.8755832912 0.5622083544
4	0.34 0.86	0.6510683761 0.3489316239
5	0 0.54 0.91	0.5652007082 0.4860117674 0.2313878782
6	0.24 0.66 0.93	0.4694282398 0.3554559718 0.1751157884
7	0 0.40 0.74 0.95	0.4087667652 0.3816431469 0.2855469914 0.1284264790
9	0 0.32 0.61 0.84 0.97	0.3274777490 0.3073797883 0.2682840654 0.1834544443 0.0771428275

Т а б л и ц а X V

Квадратура, выраженная через краевые значения (см. стр. 423)

Промежуток [0, 1]. Численные коэффициенты D_k^n умножаются на производные в конечных точках (симметризованные для производных четного порядка и антисимметризованные для производных нечетного порядка), начиная с низшего порядка, например, третье приближение будет

$$A_3 = \frac{1}{2} [f(0) + f(1)] + \frac{3}{28} [f'(0) - f'(1)] + \frac{1}{84} [f''(0) + f''(1)] + \frac{1}{1680} [f'''(0) - f'''(1)]$$

$n = 0:$	$\frac{1}{2}$
$n = 1:$	$\frac{6, 1}{12}$
$n = 2:$	$\frac{60, 12, 1}{120}$
$n = 3:$	$\frac{840, 180, 20, 1}{1680}$
$n = 4:$	$\frac{15120, 3360, 420, 30, 1}{30240}$
$n = 5:$	$\frac{332640, 75600, 10080, 840, 42, 1}{665280}$
$n = 6:$	$\frac{8648640, 1995840, 277200, 25200, 1512, 56, 1}{17297280}$
$n = 7:$	$\frac{259459200, 60540480, 8648640, 831600, 55440, 2520, 72, 1}{518918400}$
$n = 8:$	$\frac{8821612800, 2075673600, 302702400, 30270240, 2162160, 110880, 3960, 90, 1}{17643225600}$ (знаменатель: 17643225600)

ЛИТЕРАТУРА

- { 1} Курант Р. и Гильберт Д., Методы математической физики, т. I, Гостехиздат, М.—Л., 1951.
 - { 1a} Смирнов В. И., Курс высшей математики, т. I—V, Физматгиз.
 - { 1b} Крылов А. Н., Лекции по приближенным вычислениям, Гостехиздат, 1954.
 - { 2} Doherty R. E. and Keller E. G., Mathematics of Modern Engineering, Vols. I and II (John Wiley & Sons, Inc., New York, 1936 and 1942).
 - { 3} Dwyer P. S., Linear Computations (John Wiley & Sons, Inc., New York, 1951).
 - { 4} Hartree D. R., Numerical Analysis (Clarendon Press, Oxford, 1952).
 - { 5} Хаусхолдер А., Основы численного анализа, ИЛ, 1956.
 - { 6} Jaeger J. C., Applied Mathematics (Clarendon Press, Oxford, 1951).
 - { 7} Карман Т. и Био М., Математические методы в инженерном деле, Гостехиздат, 1948.
 - { 7a} Микеладзе Ш. Е., Численные методы математического анализа, Гостехиздат, М., 1953.
 - { 8} Милн В., Численный анализ, ИЛ, М., 1951.
 - { 9} Murnaghan T. D., Introduction to Applied Mathematics (John Wiley & Sons, Inc., New York, 1948).
 - { 10} Pipes L. A., Applied Mathematics for Engineers and Physicists (McGraw-Hill book Company, Inc., New York, 1946).
 - { 11} Скарборо И., Численные методы математического анализа, ГТТИ, М.—Л., 1934.
 - { 12} Schelkunoff S. A., Applied Mathematics for Engineers and Scientists (D. Van Nostrand Company, Inc., New York, 1948).
 - { 13} Smith L. P., Mathematical Methods for Scientists and Engineers (Prentice-Hall, Inc., Englewood Cliffs, N. Y., 1953).
 - { 14} Уиттекер Э. и Робинсон Г., Математическая обработка результатов наблюдений, ГТТИ, М.—Л., 1933.
 - { 15} Уиттекер Э. и Ватсон Г., Курс современного анализа, ч. I, ГТТИ, М.—Л., 1933.
 - { 16} Фаддеева В. Н., Вычислительные методы линейной алгебры, Гостехиздат, М.—Л., 1950.
 - { 16a} Фаддеев Д. К. и Фаддеева В. Н., Вычислительные методы в линейной алгебре, Физматгиз, 1960,
 - { 17} Березин И. С. и Жидков Н. П., Методы вычислений, ч. I и II, Физматгиз, М., 1959.
 - { 18} Демидович Б. П. и Марон И. А., Основы вычислительной математики, Физматгиз, 1960.
-

АЛФАВИТНЫЙ УКАЗАТЕЛЬ

- Абель* 23
Айткена метод 318
Алгебра матричная, законы 73, 79
 — —, фундаментальное правило 92
 — обыкновенная, законы 73, 79
Алгебраические уравнения высших степеней 39
 — — линейные 70
 — — —, однородная система 74
 — —, основная теорема алгебры 23
 — — с вещественными коэффициентами 39
 — — — комплексными коэффициентами 59
Алгебраическое деление 30
 — — двух полиномов 42
 — уравнение кубическое 24
 — — —, сведение к квадратному 26
 — —, метод Рауса испытания на устойчивость 61
 — — четвертой степени 36
Амплитуды комплексных корней 53
 — — —, таблица 497
Анализ полный собственных значений матрицы, численный пример 81, 85—88
 — устойчивости алгебраического уравнения, численные примеры 61, 62
Аналитического продолжения метод 436
Аналитическое разложение с помощью обратных радиусов, численный пример 435, 439
Аппроксимация равномерная 223
 — функции 361
 — —, геометрическая интерпретация 361
Архимед 312
Асимптотические разложения 458

Бернулли Д. 40
Бернулли И. 257
Бернулли метод 43, 189
Бессель 313

Бесселя формула 317, 386
 — —, численный пример 317
Биортогональное отношение сопряженных векторов 113
Борель Э. 313, 351

Вейерштрасс 66, 351
Вектор 67, 70, 88
 — ковариантный 116
 — контравариантный 116
 — n -мерного пространства 70
 — пробный 195
Векторы биортогональные 89
 — — нормированные 89
 — коллинеарные 70
 — ортогональные 89
 —, скалярное умножение 116
 —, — — в косоугольной системе координат 116
 — сопряженные 113
 — —, биортогональное отношение 113
Веса гауссовой квадратуры 515
Виета Ф. 65
Внутреннее произведение см. Скалярное произведение
Возмущение 158
Возмущений метод 158, 159
 — —, примеры приложения 159, 161, 162
Волна квадратная 232
Время запоминания 269
Вторых сумм метод 200
Вход 264, 265
Вход—выход, эмпирическое определение 268
Выделение показательных функций 280
Выход 264, 265

Галуа 23
Гамильтон 78
Гамильтона — Кели теорема 78
 — — тождество 77, 79, 80

- Ган* 313
 Гармонический анализ равноотстоящих данных 250
 — ряд, преобразование в ряд, расположенный по полиномам Чебышева 509
 — —, таблица преобразований в ряд, расположенный по полиномам Чебышева 509
Гаусс 23, 134, 312, 368, 434
 Гаусса метод исключения 134
 — — —, численный пример 135
 — — квадратур 398, 402
 — — — для случая округленных узлов 407
 — — —, нули и веса 515
 — — —, погрешность 404
 — — —, применение в технике 411
 — — —, — двукратных корней (численный пример) 409
 — — —, численный пример 401
 — уравнение 368
 — формула оценки погрешности квадратуры 404, 405
 Гершгорна теорема 193
 Гиббса колебания 235
Гильберт 434
 Гипергеометрический ряд 368
 Главные оси матрицы 76
 — — поверхности второго порядка 100, 101
 Горнера схема 29, 30, 35
Грегори 313
 Грегори — Ньютона интерполяционная формула 314
 Грина тождество 364

Декарт Р. 97
 Дельта-функция 233
 —, разложение в тригонометрический ряд 233
 Диагональная матрица 91, 109, 132
Дирихле 218, 220
 — условия 219
 — ядро 220
 Дифференциальное уравнение Гаусса 368
 Дифференциальный оператор самосопряженный 365
 — — Штурма — Лиувилля 366
 Дифференцирование табулированной функции 319

 Единичная матрица 133
 Единичный импульс 266

 Задача взвешенных моментов 282
 — «обращения» матрицы 80
 Задачи собственных значений 424

 Избранных точек метод 493
 Импульс единичный 266
 Импульсная реакция 266, 291
 Инвариантность определителя матрицы 149
 — транспонирования матриц 128, 131
 — уравнений матричной алгебры 126, 131
 Интеграл Фурье 257
 Интегральная показательная функция 439
 — формула Коши 442
 Интегральное преобразование 416
 Интегральный синус 275
 Интерполяционная формула Грегори — Ньютона 314
 — — Лагранжа 397
 Интерполяционные формулы для неравноотстоящих значений аргумента 318
 Интерполяционный анализ фильтра 273
 Интерполяция косинусами 247
 — ортогональными полиномами 372
 — полиномами Чебышева 254
 — — — равноотстоящих данных 353, 354, 355
 — полиномиальная в целом 349
 — простыми разностями 313
 — синусами 244
 — тригонометрическая 238
 — —, точность 251
 — центральными разностями 315
 Итерационное решение обширных линейных систем 201, 205

 Квадратная волна 232
 Квадратура, выраженная через крайние значения 420, 516
 Квадратура Гаусса см. Гаусса метод квадратур
 — механическая 381
 — при помощи дифференцирования 417
 — — — планиметров 381
 —, содержащая показательные функции 416
 Квадратурная формула, вычисление коэффициентов 406
Кели 66, 68, 78
Кеплер 97
 Кодиагональ 153
 Кодиагональная система 153

- Кодовые числа 333
 Колебания Гиббса 235
 — —, сглаживание с помощью σ -множителей 235
 Комплексное число, извлечение корня 490
 — —, подстановка в полином 33
 Компоненты вектора ковариантные, контравариантные 116
 Конические сечения 97
 Координаты вектора 67
 — — n -мерного пространства 70
 — тензора 68
 Корень, наибольший по модулю 43, 47, 49
 — неустойчивости 61
 — —, локализация 61
 Корни полинома, близкие к мнимой оси 56
 — — кратные 58
 — —, локализация вдоль окружности наибольшего радиуса 47, 53
Коши 24
 — интегральная формула 442
 Коэффициенты гибкие 444
 — жесткие 444
 — квадратурной формулы с произвольными узлами 406
 — полиномов Лагерра 305, 506
 — — Лежандра 295, 502
 — смещенных полиномов Чебышева 297, 504, 505
 — Фурье 219
 — —, затухание 232
 Краевые условия сопряженные 365
 Кронекера символ 397
 Кубические уравнения 24, 33

Лагерра полиномы 303, 371
 — функции 304
 — — нормированные 506
Лагранж 23, 183, 218, 257
 Лагранжа интерполяционная формула 397
Лаплас 343
 Лапласа преобразование 288
Лежандр 312, 434
 Лежандра полиномы 294, 370
Лейбниц 65
Лиувиль 366

Маклорен 433
 Матрица 68
 — большая 155
 —, главные оси 76
 — диагональная 91, 109, 132
 Матрица единичная 133
 — —, геометрическое представление 133
 —, инвариантность определителя 149
 —, инварианты 126, 128, 129, 131
 — калиброванная 194
 — комплексная, обращение 151
 — —, переход к вещественной 152
 —, метод обращения 138
 — неполноосная 124
 — обратная 80, 133
 —, обращение подразделением на блоки 155
 — обширная 152, 155
 — полноосная 124
 —, полный анализ собственных значений (численный пример) 81, 85—88
 — положительно определенная 147, 201
 —, последовательная ортогонализация 137
 — почти треугольная 161
 —, преобразование к главным осям 105, 109
 —, разложение в ряд по параметру возмущения 160
 — рекуррентная 187
 — с выделенными по величине диагональными элементами 109
 — самосопряженная 104
 — симметрическая 100
 —, случай комплексных собственных значений 121
 —, — кратных собственных значений 109
 —, собственные векторы 76
 —, — значения 75
 — транспонированная 82
 — треугольная 141
 — —, обращение 144
 —, триангуляризация 150
 —, характеристические значения 75
 —, характеристический полином 75
 —, характеристическое уравнение 75
 — эрмита 104
 — —, наименьшее собственное значение 211
 — —, свойство собственных значений 104
 Матрицы, алгебраическое тождество 78
 — коммутативные, некоммутативные 132
 — преобразований, таблица 497
 —, сложение 69
 —, транспонирование произведения 92
 —, умножение 70—72, 133
 —, условие коммутативности 132

- Матричная алгебра, законы** 73, 79
 — —, отличие от обычной 79
 — —, фундаментальное правило 92
 — запись системы уравнений 91
Матричные равенства, инвариантность относительно ортогональных преобразований 128
 — —, — — произвольных линейных преобразований 131
Матричный множитель, симметрические свойства 248
 — полином 184
Метод см. соответствующее название
Множители затухания 232
Множитель веса $1/2$ 248
Момент взвешенный 282
 — функции 294
Моментов метод 40, 59
Моменты симметрические корней 41

Наименьших квадратов метод 142, 170, 223, 312
 — — —, геометрическая интерпретация 360
 — — —, основной принцип 321
 — — —, полиномы 347
Нарастающая полоса 30
Невилль 318
Невязка 322
Неподвижная полоса 30
Неясная зона 341
 n -мерная поверхность второго порядка 101
 n -мерное пространство 98
 — —, векторы линейно независимые 111
 — —, косоугольная система координат 111
 — —, матрица преобразования координат 106
 — —, ортогональное преобразование координат 108
 — —, преобразование координат 105
 n -мерный эллипсоид 98
Нормальные уравнения 322, 349
Нули гауссовой квадратуры 515
Ньютон 313
Ньютона метод 27, 35, 37
 — —, итерационная схема 28
 — —, улучшение сходимости 28
 — —, численный пример 29

Обратная матрица 80, 133
Обращение комплексной матрицы 151
 — матрицы 80, 81, 133
 — — методом возмущений 159

Обращение матрицы подразделением на блоки 155, 157
 — —, численный пример 136
 — преобразования Лапласа 292
 — — — с помощью полиномов Лежандра 293
 — — — — — Чебышева 296
 — — — — — рядов Фурье 297
 — — — — — функций Лагерра 300
 — собственных значений 213
 — — —, таблица 500
 — треугольной матрицы, численный алгоритм 145, 146
Обширная система 155
Операции над матрицами 69, 71, 72, 83, 91, 94
 — — —, геометрическая интерпретация 96
Ортогонализация векторов приближенная 142
 — линейной системы 175
 — матрицы 137
 — — путем вращения системы координат 175
 — —, численный пример 147
Ортогональная система векторов 102
Ортогональное разложение функций 363
 — семейство функций 224
Ортогональности условие 223
Ортогональность векторов 89
 — взвешенная 367
 — полиномов, общее условие 400
 — собственных векторов, алгебраическое доказательство 90
 — функций Фурье 225
Ортогональные полиномы, трехчленное рекуррентное соотношение 374
 — функции 362
Ортонормированное семейство функций 224
Основная теорема алгебры 23
Остаточное испытание решения линейной системы 210
Остаточный вектор 162, 210
Относительная ошибка решения линейной системы 208
Отображение «типа С» 264
Оценка наибольшего по модулю корня алгебраического уравнения 47
Ошибка средняя 223, 322
 — тригонометрической интерполяции 251

Периодичности скрытые 196, 276
Пиковое значение 200

- Пиковых значений метод 47
Пифагор 218
p, q-алгоритм 186, 188
 Поверхность второго порядка, преобразование к главным осям в косоугольной системе 117
 Подвижная полоса 30
 Подходящие кривые, таблица 510
 Полином матричный 184
 Полиномиальная интерполяция в целом 349
 — — равноотстоящих данных 353—355, 359
 — —, сходимость 355
 Полиномы, алгебраическое деление 41, 61
 — канонические 463
 — Лагерра 303, 371
 — —, таблица 506
 — Лежандра 294, 370
 — —, таблица 502
 — метода наименьших квадратов 347
 — ультрасферические 370
 — —, случай $\gamma = \infty$ 371
 — Чебышева 190, 193, 255, 370, 503
 — — второго рода 192, 371
 — —, рекуррентная формула 191, 192, 449
 — — смещенные 192
 — — — второго рода 192
 — —, таблицы 503—505
 — Якоби 370
 Полная ортогональная система функций 363
 Помехи 160, 163, 166, 330, 337
 —, их влияние на решение обширных систем 179
 —, локальный прием их исключения 336
 Преобразование координат 105, 109
 — Лапласа 288
 — — и анализ цепей 290
 — —, интерполяционная формула 306, 308
 — —, связь с преобразованием Фурье 289
 — матрицы к главным осям 105, 109
 — обратное 54
 — обратными радиусами 47, 53, 61
 — Фурье 260
 — —, интерполяционная формула 272
 Приближение полиномами высших степеней 393
 Проблема собственных значений, геометрическая интерпретация 96, 101
 Произведение матриц 71, 72
 — числа на вектор 70
 Радиус-вектор 100
 Разложения гибкие 444
 — жесткие 444
 — по ортогональным полиномам 445
 Разности отрицательного порядка 331
 — простые 319
 — разделенные 318
 — центральные 316, 319
 Разностные коэффициенты 313
 Разыскание скрытых периодичностей 196
Раус 61
 Реакция установившаяся 265
 — частотная 265, 291
Релей 434
 Релея — Ритца метод 427
 Решение алгебраического уравнения кубического с помощью полиномов Чебышева 26
 — — —, численный пример 26, 33
 — — — четвертой степени 36, 37, 39
Рунге 244, 313, 351, 359
 Ряд гипергеометрический 368
 — Тейлора 313
 — —, сходимость 441
 — Фурье 220
 — —, дифференцирование 230
 Самосопряженная матрица 104
 — совокупность векторов 113
 Самосопряженный дифференциальный оператор 365
 Сглаживание в малом 336
 — σ -множителями 235, 237
 — эмпирических формул четвертыми разностями, численный пример 323, 325
 Сглаживания параметр 399
 σ -множители 50, 235
 —, общий характер сглаживания 237
 —, сглаживание колебаний Гиббса 235
 σ -сглаживание 50
 Сигнал 264
Сильвестр 66
 Символ Кронекера 397
 Симметризация матрицы 202
 Симметрическая матрица 100
 — —, свойства собственных значений 103
 — функция 41
 Симметрические моменты корней 41
 Симпсона формула 384
 — — с концевой поправкой 412, 414, 415
 — —, точность 385
 — —, численный пример 391, 393
 Синтез Фурье 261

- Система векторов ортогональная и нормальная 89
 — координат 110, 117
 — — весьма косоугольная 143
 — — косоугольная 110, 111, 113, 114, 115
 — — прямоугольная 112, 113
 — линейных уравнений весьма косоугольная 143
 — — —, выправление переопределенности и несовместности с помощью метода наименьших квадратов 169, 170
 — — — кодиagonalная 153
 — — —, косоугольность 173
 — — —, критерий физической надежности 180
 — — —, матричная запись 91
 — — —, недостаточная определенность 166
 — — — обширная 155
 — — — —, итерационное решение 201
 — — —, определенность 166
 — — —, ортогонализация 175
 — — —, остаточное испытание решения 210
 — — —, относительная ошибка решения 208
 — — —, отыскание наилучшего решения 170
 — — —, переопределенность 170
 — — —, решение методом возмущений 162
 — — —, совместность 164, 165
 — — —, степень определенности (численный пример) 166
 Скалярное произведение векторов 88, 116
 — — функций 361
 Скрытых периодичностей свойство 47
 Сложение матриц 69
 Собственное значение матрицы 75
 — — —, отличное от наибольшего 214
 Собственные значения 75
 — —, геометрическая интерпретация 101
 — — комплексные 121
 — — кратные 109
 — —, основное уравнение 101
 — —, полный анализ 205
 Собственный вектор 76, 166
 Совместность линейных систем 164, 166
 Совокупность векторов самосопряженная 113
 Составляющие вектора n -мерного пространства 70
 — тензора 68
 Спектроскопический алгоритм 197, 199
 — метод 197
 Степенные суммы 41, 43
 — — взвешенные 45
 — —, типовые значения 47
 — —, свойства 43, 44
 Степенных сумм метод 40, 59
 Стирлинг 313
 Стирлинга разложение 257
 — —, остаточный член 257
 — формула 258, 317, 385
 — —, численный пример 317
 Строка таблицы полная 316
 — — половинная 316
 Сумма векторов 70
 — матриц 69
 Схема Горнера 29, 35
 Таблица разностей 314
 — — для эмпирических функций 320
 — —, неудобства 319
 — центральных разностей 316
 Тарталья 23
 τ -метод 457
 —, обобщение 493
 —, оценка погрешности 463
 —, примеры 467
 Тейлор 433
 Тейлора ряд 313
 Телескопический сдвиг степенного ряда 451, 453
 Тензор 68
 — второго ранга 68
 Теорема см. соответствующее название
 Техника подвижной полосы 30, 32
 — — —, расположение с обратной связью 31
 Транспонированная матрица 82
 Трапедий правило 341, 386, 387
 — формула 381
 — — с концевой поправкой 387
 — —, точность 386
 — —, численный пример 390, 392
 Триангуляризация матрицы 150
 Тригонометрический ряд, распространение на неинтегрируемые функции 234
 Тригонометрической интерполяции метод 238
 Уменьшения случайного разброса метод 324
 Умножение матриц 71, 72
 Уравнение см. соответствующее название

Уравнение поверхности второго порядка в косоугольной системе координат 117

Условие см. соответствующее название

Устойчивость 60

—, достаточный критерий 62

Уэддла формула 396

Фейер 219, 221, 313

Фейера метод 221

Феррари 23

Фильтр 273

— идеальный 275

Фокусирующее действие функции 220

Формула Бесселя 317

— Симпсона 384

— — с концевой поправкой 412

— —, точность 385

— Стирлинга 317

— трапеций 381

— — с концевой поправкой 387

— —, точность 386

— Уэддла 396

Формулы оценки погрешности квадратуры Гаусса 404, 405

Фредгольм 66, 183, 434

Фробениус 66

Функции Лагерра 304

— — нормированные 506

— —, таблица 506

— фундаментальные, соответствующие таблицы центральных разностей 317

— Фурье 225

Функция абсолютно интегрируемая 221

— входа 264

— выхода 264

— гипергеометрическая 368

— нечетная 227

— передаточная 265

— симметрическая 41

— $S(t)$ 49

— табулированная, дифференцирование 319

— целая 435

— четная 227

— эмпирическая 323

Функция, эффективное приближение 362

Фурье 218

— интеграл 257

— коэффициенты 219

— преобразование 260

— ряд 220

— синтез 261

— функции 225

Характеристические значения 75

Характеристический полином матрицы 75

Характеристическое уравнение матрицы 75

Хевисайд 290

Хевисайда теорема разложения 291

Центральная поверхность второго порядка 100

Цепь электрическая 290

С-итерация 209

Частные суммы 222

Частота граничная 338

— —, эмпирическое определение 341

Чебышев П. Л. 190, 435

Чебышева полиномы 190, 370, 503

— — второго рода 371

Штурма — Лиувилля дифференциальный оператор 366

Эйлер 23, 218, 257

Элементы матрицы 69

Эллипсоид n -мерный, уравнение 98, 100

Эмпирическая функция, вторая производная 332

— —, дифференцирование 327

— —, — с помощью интегрирования 330

— —, сглаживание 323

— —, — в целом 336

Эрмитова матрица 104

— —, наименьшее собственное значение 211

Ядро Дирихле 220

Якоби полиномы 370